THE ROLE OF THE AUDITORY SIGNAL IN AUDITORY-VISUAL INTEGRATION

CAPSTONE PROJECT

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DOCTOR OF AUDIOLOGY

IN THE GRADUATE SCHOOL OF THE OHIO STATE UNIVERSITY

BY

NATALIE M. FELEPPELLE

*****

THE OHIO STATE UNIVERSITY
2008

CAPSTONE COMMITTEE
JANET WEISENBERGER, PH.D., ADVISOR
ROBERT FOX, PH.D.
CHRISTINA ROUP, PH.D.
GAIL WHITELAW, PHD.

APPROVED BY:

_____
ADVISOR

ABSTRACT

Although numerous studies have documented the existence of auditory-visual integration for both perfectly intelligible and compromised speech, research has yet to determine if it is these compromised situations that optimize this process. There is compelling evidence suggesting that listeners benefit from the addition of visual cues when auditory information is compromised in some way. However, studies in multimodal perception document a large degree of variability in the amount of benefit listeners receive from auditory-visual integration. The present study examined the role of the auditory signal in auditory-visual integration in order to explore how characteristics of the auditory signal might influence and change this process. The present study addressed several specific questions including: *Does the amount of available information in the auditory signal affect auditory-visual integration ability? How does the integration process change as auditory information available in the speech signal is altered? Is the amount of information available in the auditory signal a contributing factor to the large degree of variability noted in the amount of benefit listeners receive from auditory-visual integration?*

The present study examined if the amount of information available in the auditory signal affected the auditory-visual integration process. Listeners were presented with degraded speech stimuli containing different amounts of information and their speech perception abilities in a number of conditions were measured. The stimuli were degraded using 2, 4, 6, and 8 bandpass filter channels. The temporal fine structure was then

removed from the speech syllable stimuli and was replaced with narrowband noise, modulated by the temporal envelope retained in the filtered speech waveform. Performance was examined in four different conditions: auditory-only (A), visual-only (V), congruent auditory + visual (AV) and discrepant auditory + visual (AV).

Results of this study demonstrated that degrading auditory stimuli, regardless to what degree, decreases overall speech perception ability in the A and AV conditions. Results also revealed a decrease in the amount of auditory-visual integration listeners achieved for discrepant stimuli when compared to the amount of McGurk-type integration they are able to achieve when the auditory signal is perfect. Although overall integration ability is decreased for degraded stimuli, removing information from the auditory signal does not inhibit auditory-visual integration ability. Surprisingly, the amount of auditory-visual integration did not change across the majority of the different levels of auditory signal degrading. Finally, there do appear to be difference in the type of stimulus materials used and the way that listeners integrate auditory and visual speech.

The present study has provided some important insight to the process of auditory-visual integration. It is clear from the results of this study that removing information from the auditory signal does affect auditory-visual integration, but does not inhibit this ability. However, as the amount of signal degrading is altered only one level of degrading showed slightly significant differences in the amount of auditory-visual integration that was achieved at that level. These results suggest that the amount of information available in the auditory signal may play a slight role in the variability noted in the amount of benefit listeners receive from auditory-visual integration.

DEDICATION

I would like to dedicate this capstone project to my advisor, Janet Weisenberger, Ph.D., who has guided me throughout this project. She has taught me so much about research through work in her research lab, directed study, and personal communication and I will forever have a great respect and appreciation for the science behind practice. I would to thank her for her commitment to my growth and development as a person and professional. She has given me strength and confidence which have helped me to succeed throughout my professional education and career. She is a remarkable advisor, educator, researcher and friend and a true inspiration.

ACKNOWLEDGEMENTS

I would like to thank my capstone committee Robert Fox, Ph.D., Christina Roup, Ph.D., and Gail Whitelaw, Ph.D., for their guidance and assistance in all aspects of this project. I would like to thank my subjects for their participation in this study; for without them this research would not be possible. Finally, I would like to thank my family for their support and encouragement.

VITA

DECEMBER 15, 1980. …….. . . . . BORN IN CLEVELAND, OHIO.

AUGUST 2001. . ……... … . . ASSOCIATE OF APPLIED SCIENCE, THE OHIO STATE UNIVERSITY AGRICULTURAL TECHNICAL INSTITUTE.

JUNE 2004……..…….BACHELOR OF ARTS WITH DISTINCTION, THE OHIO STATE UNIVERSITY.

2004-2008. . . . . ……... . . . . GRADUATE RESEARCH ASSISTANT, THE OHIO STATE UNIVERSITY.

FIELDS OF STUDY

MAJOR FIELD: AUDIOLOGY

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Studies of the multimodal perception of speech have offered insights into the different resources available for successful speech perception. While speech perception is often thought to be a unimodal process (i.e., using a single sense), it is in fact a multimodal process that integrates both auditory and visual inputs. The auditory signal is generally considered the dominant modality for understanding speech in normal-hearing individuals. However, research suggests that auditory-visual integration is a natural part of speech perception and individuals use visual information during communication even when the auditory signal is highly intelligible (McGurk & MacDonald, 1976).

What is less obvious is that visual information is also an important component of successful speech perception. Listeners have the ability to integrate auditory and visual cues when the auditory signal contains very little information and receive significant benefits from this integration. Visual cues are especially relevant in situations where the auditory signal is compromised in some way (e.g., loss of hearing, listening in background noise, or listening to unfamiliar dialects). Visual cues enhance features of the auditory signal and supplement missing auditory information and aid in the detection and perception of speech in these conditions (Grant, 2000).

Auditory-visual integration can aid listeners during speech perception in high levels of background noise. Grant and Braida (1991) found that listeners achieved significantly higher levels of speech perception in poor signal-to-noise ratios by adding visual cues were added, compared to auditory speech in noise alone. It has also been suggested that adding visual cues to speech in noise may improve speech perception performance by at least five to ten percent. Further, Grant, Walden and Seitz (1998) suggested that the addition of visual cues to compromised auditory stimuli improved the listener's ability to use syntactic and linguistic information to understand speech.

Degraded auditory signals can become highly intelligible when presented with visual cues during speech perception. Through the integration of auditory and visual cues listeners can rely less on damaged auditory cues, because visual cues can significantly enhance perception abilities of the compromised speech information (Shannon et al., 1998; Shannon et al., 1995; Munhall et al, 2004; Grant 2000). Grant and Seitz (1998) observed auditory-visual integration abilities of listeners using congruent and discrepant (i.e., McGurk stimuli) auditory-visual nonsense syllable and sentence recognition tasks. They showed that even a highly degraded auditory signal (i.e., single formant sine wave speech using only the frequency transitions of the fundamental frequency F0) presented with visual cues was intelligible to listeners (Grant & Seitz, 1998).

Individuals with certain types of hearing loss experience unique challenges during communication. Due to peripheral and central auditory system damage these listeners receive degraded auditory information, which can be further compromised in noisy or difficult listening situations. Listeners with hearing loss are also able to integrate auditory and visual cues and receive benefits from this integration. Grant, Walden, and

Seitz (1998) suggest auditory-visual communication may address some of the hearing acuity issues experienced by older adults with hearing loss, particularly in consonant identification and sentence recognition.

Although there is compelling evidence that suggests listeners benefit from the addition of visual cues when auditory information is compromised in some way, there is a large degree of variability noted in the amount of benefit listeners receive from auditory-visual integration. Researchers have hypothesized several factors that may contribute to the variability in auditory-visual integration performance seen in and across listeners. These factors include: the measurement technique used for reporting auditory-visual integration or benefit, individual characteristics of the listener or talker, and the acoustic properties of the speech signal.

While much of the current research has focused on addressing whether auditory-visual integration occurs, less attention has been placed on whether integration ability and the characteristics of integration change as a result of external sources of variability such as different talkers and auditory inputs. One approach to understanding the differences in auditory-visual integration abilities across listeners is to examine those possible sources of variability. Understanding the root of this variability may aid in determining if these abilities can be optimized.

The present study examined if the amount of information available in the auditory signal affected the auditory-visual integration process. Listeners were presented with degraded speech stimuli containing different amounts of information and their speech perception abilities in a number of conditions were measured. The stimuli were degraded using 2, 4, 6, and 8 bandpass filter channels. The temporal fine structure was then

removed from the speech syllable stimuli and was replaced with narrowband noise, modulated by the temporal envelope retained in the filtered speech waveform. Performance was examined in four different conditions: auditory-only (A), visual-only (V), congruent auditory + visual (AV) and discrepant auditory + visual (AV). Performance was examined in four different conditions: auditory-only (A), visual-only (V), congruent auditory + visual (AV) and discrepant auditory + visual (AV) (e.g., visual /gat/ simultaneously presented with auditory /bat/).

The purpose of the present study was to examine several specific questions about this multimodal process including: *Does the amount of available information in the auditory signal affect auditory-visual integration ability? How does the integration process change as auditory information available in the speech signal is altered? Is the amount of information available in the auditory signal a contributing factor to the large degree of variability noted in the amount of benefit listeners receive from auditory-visual integration?*

The results of the present study should provide a clearer understanding of the role the auditory signal plays in auditory-visual integration and should also provide further insight into the factors that are responsible for the variability noted in the amount of benefit listeners receive from auditory-visual integration. An enhanced understanding of acoustic stimulus factors will also inform clinical recommendations for the signal processing strategies of cochlear implants and hearing aids. Finally, the results have implications for the design of aural rehabilitation programs in suggesting how listeners can maximize auditory-visual integration.

CHAPTER 2

LITERATURE REVIEW

AUDITION AND SPEECH PERCEPTION

To understand auditory-visual integration it is necessary to understand the information provided in the auditory and visual signal important for speech perception. The auditory signal contains spectral and temporal cues that convey the information necessary for listeners to perceive speech. The spectral content of the speech signal depends on largely on the laryngeal tone and on the filtering effects of the throat, mouth, and nasal cavities, which are continuously modified during different speech sound productions. The laryngeal tone, to some extent, will impact the fundamental frequency of a sound that is produced. The fundamental frequency is the most concentrated frequency band and it usually contains the most energy. Resonances in the vocal tract reinforce the sounds produced within a particular frequency band. Several of these frequency bands (resonances) are known as frequency formants and together with the fundamental frequency provide information about vowels and diphthongs. The fundamental frequency of vowel sounds are typically associated with small highly concentrated frequency bands of only a few hertz. The fundamental frequencies of vowels usually lie in the low- and mid- frequency regions, below 1000 Hz. Vowel resonances tend to have a bandwidth of about 200 Hz, which is still highly concentrated

although they contain less energy, and these formants lie higher in the spectrum than do their fundamental counterparts. Vowel sounds are classified by where they are made in the vocal tract or mouth. Most linguists recognize between 17 and 21 distinct vowel sounds, all created by air passing over the vocal chords and shaped by the mouth.

Consonants are transient or transitional sounds that are distinct from vowels, and more sustained in character. Their frequency content is not always clearly distinguishable from vowels because it is significantly weaker and occupies a larger range and transition of frequencies. However, formant transitions occurring in the upper-mid- to high-frequency regions typically above 1500 Hz, convey most consonant information. Consonants come in several varieties characterized by which parts of the mouth (lips, tongue, teeth, palate) and throat (soft palate, uvula, larynx, pharynx) are used and how the expelled air is manipulated by them. Place of articulation information, the part of the mouth or throat that modifies air flow and forms the specific sound, is typically conveyed by these high-frequency formant transitions across the speech spectrum. Manner of articulation information, i.e., the way that the mouth or vocal tract regulates airflow, is also conveyed by subtle spectral cues, like the small burst of initial low frequency energy produced in a nasal sound, for example.

In the speech signal, maximum energy is typically found in the 250 Hz octave region with a little less energy spreading into the 500 Hz region. These lower-frequency bands, as previously mentioned, roughly correspond to vowel sounds, whereas the higher-frequency consonant sounds congregate around the 2 and 4 kHz regions. Some exceptions include, the /i/ vowel sound with most of its energy above 2 kHz and the /s/ sound and its many sibilants occurring in the 4 and 8 kHz bands (Ladefoged, 1996).

As a rule in western languages, vowels provide the power of speech (high energy at low frequencies) while the less powerful consonants provide speech intelligibility. It is interesting to note that there is around a 27 to 28 dB difference in the power between the strongest vowel phoneme and the weakest consonant sound, which corresponds to a range of around 500:1 (Ladefoged, 1996). Studies demonstrate that the 125 and 250 Hz regions provide little speech intelligibility, and the 500 Hz band only provides around 12% speech intelligibility. However, these lower frequencies are important for talker recognition and the overall rhythm of the speech. The 2000 Hz band contributes the most to speech intelligibility. Together with the 4 kHz band, they provide some 57% of overall syllable and word discrimination ability (Grant et al., 2006; Grant and Greenburg, 2001)

Temporal cues carry a surprising amount of additional information that can help listeners to discern the contents of the speech signal. Temporal cues convey timing and amplitude envelope (energy) information, which can vary subtly between different sounds. Manner of articulation information, which is identified by temporal duration, provides information about the slight differences in phoneme sounds. The length of a sound will distinguish areas of concentrated spectral energy and can be used to distinguish different groups of visemes and phonemes and lead to the identification of certain sounds. Voicing information is conveyed by changes in energy intensity, and differences in voicing onset time and energy patterns also distinguish different sounds. The temporal cues found in speech are sufficient enough to convey almost all of the information on manner and voicing and aid in consonant identification and speech intelligibility (Shannon, Zeng, Kamath, Wygonski, Ekelid, 1995).

VISION AND SPEECH PERCEPTION

During natural face-to-face conversation, visual cues such as head movements, lip and mouth movements, and facial animation (i.e., eye brow or facial movement) provide information to listeners that aids in the understanding of a spoken message. These nonverbal visual cues convey prosodic characteristics (e.g., the rhythmic aspect of language and suprasegmental features such as pitch, stress, and juncture) and acoustic properties (e.g., fundamental frequency changes, amplitude envelope, and place of articulation) of the speech signal. Studies of visual prosody suggest that natural head movements are related to prosodic characteristics of speech such as syllable or word stress and prominence (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). Research on facial movement shows that facial animation helps listeners to determine the emphatic stress of a message and to discriminate among statements, questions, and directives (Nicholsen, Baum, Cuddy, & Munhall, 2002; Srinivasan & Massaro, 2002).

Visual cues also provide significant information about the acoustic properties of speech and aid in speech awareness. Head gestures can provide cues about the spectral content in the speech signal, mainly frequency and amplitude patterns of a talker's voice (Hadar, Steiner, Grant, & Rose, 1983; Munhall et al., 2004). Movements of the lips and mouth during speech mimic the acoustic speech envelope and frequency transitions in the speech signal (Munhall et. al., 2004; Grant, 2001). Lip and mouth cues also provide information about the place of articulation of speech sounds, helping to distinguish groups of phonemes (Grant & Seitz, 2000). Speech sounds that have similar visual patterns are divided into phoneme groups known as visemes. Although visemes are beneficial to listeners, they allow listeners only to distinguish among groups of sounds,

rather than individual sounds within the group (Jackson, 1988). For example, the viseme group /b, p, m/ consists of bilabial consonants, which are all produced by closure of the lips, making it difficult for a listener to distinguish among these sounds using visual cues alone. However, when speech sounds have similar visual characteristics, auditory cues allow listeners to distinguish between individual phonemic sounds. Although visual cues alone are rather ambiguous, they provide cues that are redundant to the acoustic properties of speech and enhance the speech signal, improving auditory intelligibility.

Observing visual cues during face-to-face communication influences what we hear. A study by McGurk and MacDonald (1976) demonstrated that visual information plays a role in the perception of speech, even when auditory intelligibility is high. They presented listeners with a video of a talker producing an auditory syllable /bɑ/ simultaneously with a visual syllable /gɑ/, and listeners reported perceiving the syllable /dɑ/ or /thɑ/, a fusion of the places of articulation of the auditory and visual speech stimuli. However, when listeners were presented only with the auditory syllable (and no visual input) they correctly perceived the syllable /bɑ/. The inability to ignore incongruent visual cues during speech recognition has been referred to as the "McGurk effect." These findings suggest that auditory and visual inputs are integrated together and that both of these inputs can provide important information that influences the perception of speech.

AUDITORY-VISUAL INTEGRATION

Multimodal integration is the process employed by individual receivers to combine information extracted from two sensory modalities to arrive at a perceptual decision. Auditory-visual speech integration is the ability of a listener to extract incoming auditory and visual speech cues from the speech signal and then combine information from both modalities to arrive at a speech percept. The "McGurk effect" provides compelling evidence that vision influences what we hear. A growing body of research indicates that auditory-visual integration is a skill that can be measured independently of auditory and visual extraction (Grant and Seitz, 1998; Massaro, 1998; Braida, 1991). Furthermore, as a natural consequence of the integration process, listeners cannot ignore information from either modality and use both auditory and visual cues when they are available to interpret speech.

Spence (Zampini, Shore, Spence, 2003) used the term "multisensory binding" to describe how the brain uses spatiotemporal coincidence to explain how sensory stimuli are linked together to create a multimodal perceptual event. In auditory-visual speech perception, the spectral and temporal characteristics of the auditory and visual inputs must be grouped together and perceived as being coherent in order for listeners to integrate auditory and visual cues to arrive at an auditory-visual percept (Grant, Greenburg, Poeppel, & van Wassenhove, 2004; Moore, 2004).

Although there are many spectral and temporal properties in the auditory and visual signals that may be important for multimodal integration, listeners are particularly sensitive to changes in the relative timing differences of sensory inputs across modalities during auditory-visual speech perception. Studies have explored the affects of temporal

asynchrony on auditory-visual integration to determine the temporal window in which auditory and visual cues can present without affecting this process. Grant et al. (204) tested the multimodal integration abilities of listeners when auditory and visual stimuli were asynchronously presented over a wide range of auditory lead and lag durations. Results indicated that integration continued to occur over a range of asynchronous presentation of approximately -30 to +170 ms auditory lag to lead, suggesting the existence of a 200 ms duration temporal integration window. This research supports the notion that there is a temporal window within which auditory-visual integration can occur (Grant, van Wassenhove, Poeppel; Navarra, Vatakis, Zampini, Faraco, Humphreys, & Spence, 2005; Vatakis & Spence, 2005; McGurk & Mac Donald, 1976). The temporal window plays an important role in auditory-visual integration, because of the relative timing differences in the processing of auditory and visual signals. This window allows two separate signals which are processed at different speeds throughout their respective systems to be combined and perceived as a single event (Vatakis & Spence, 2006; Grant et al., 2004).

Combining different sensory cues can improve a listener's ability to reach a perceptual decision by providing the listener with more information than might be available via one sensory modality. But different behavioral or perceptual decisions may be made based on the quality and/or content of the sensory information available from the individual modalities (Ernst & Banks, 2002). As demonstrated with the McGurk effect, when sensory information from the auditory and visual modalities is incongruent, the listener's response may not reflect the auditory or visual stimulus, but some combination of the available cues. Furthermore, studies in our lab suggest that as the auditory

information is degraded during McGurk-type integration tasks, perceptual decisions may be altered based on the quality of the available information (Andrews, 2007; Anderson, 2006; ver Hulst, 2006; Yeary, 2001).

In the Ernst and Banks model, the degree to which one modality dominates the final perceptual decision is based on maximum-likelihood estimation. Ernst and Banks (2002) suggest that individual dependency on one sensory modality can be altered by changes in task demands or by variations in the quality of sensory input, termed modality salience. The principle of modality salience is evident in studies of auditory-visual integration, as is the notion that individual dependency on one sensory modality can be altered by changes in task demands or by variations in the quality of sensory input. A growing body of research demonstrates a pronounced shift in the weight assigned to visual cues as information is removed from the auditory signal, improving speech detection and recognition in auditory-visual speech when compared to auditory speech alone (Andrews, 2007; Anderson, 2006; ver Hulst, 2006;Yeary, 2001; Grant and Seitz, 2000). Grant and Seitz (2000) discovered an effect of input signal strength on speech detection thresholds, which they termed Bimodal Coherence Masking Protection (BCMP). In their study, subjects were presented with sentences in noise under three conditions: auditory alone, auditory + visual matched (the auditory and visual sentences were congruent or matched), and auditory + visual unmatched (the auditory and visual sentences were incongruent or unmatched). The primary question in this study addressed whether the weight of visual speech cues increases as the quality of the auditory signal decreases, allowing visual speech to improve speech detectability in noise. When subjects were presented with an auditory stimulus in noise, speech detection thresholds were on

average 1.6 dB more sensitive and remained protected from the masking effects of noise with the addition of congruent visual speech when compared to the auditory alone condition. However, no improvements in the detection of speech were evident when masked auditory stimuli were presented with incongruent visual speech. Results of this study support findings in multimodal integration research and suggest that visual cues provide information that is complementary to the acoustic properties available in a masked auditory signal, increasing the salience of the auditory signal and enhancing speech detectability in noise. This study also suggests that when the auditory signal is degraded the weight assigned to visual cues increases and varies along a continuum as a function of visual input magnitude.

Pilot studies in our laboratory have compared the McGurk-type auditory-visual integration abilities of listeners using degraded and undegraded auditory stimuli. Results of these studies found that in the AV condition subjects showed a heavy reliance on auditory input and relatively little reliance on the visual input, most likely due to the strength of the auditory signal and listeners' natural reliance on the auditory signal for understanding speech. Interestingly, this pattern was completely reversed in the degraded AV condition, showing a substantial decrease in auditory responses and a significant increase in visual responses. Thus, despite the relative ambiguity of the visual signal, subjects were more likely to rely on visual cues for information rather than the degraded auditory stimuli. This finding appears to reflect a change in subject behavior based on the relative strength of the auditory and visual inputs. Results of these studies suggest that listeners can learn to rely more heavily on visual information when the auditory signal is degraded in order to maintain speech recognition capabilities and supports the

notion that multimodal perceptual decisions vary as a function of sensory input strength (Andrews, 2007; Anderson, 2006; ver Hulst, 2006; Yeary, 2001).

From these studies, it is evident that individual dependence on one sensory modality can be altered by changes in task demands or by variations in the quality of individual sensory inputs. It is also evident that dominance of one modality will shift along a continuum depending on the strength of the individual sensory inputs. In auditory-visual speech perception, research suggests that while audition may be the dominant modality for speech perception in normal hearing listeners, a shift in dependency towards visual information occurs when the auditory signal is compromised. The visual cues appear to provide complementary information that may be lost in a compromised auditory signal.

It has also been suggested that integration processes may be influenced by the type of speech stimulus presented and the relative difficulty level of the task. In particular, it is possible that individuals integrate consonant and vowel syllables and words differently from connected speech (Grant, 2002). A possible reason for this may be the additional demands placed on the integration process when individuals are forced to perceive continuous speech at faster rates in a running discourse. During the perception of isolated syllables and words, individuals may have more time available to access their working memory to store and recall auditory and visual cues used in the integration process (Massaro, 1972). In contrast, during connected speech, speech segments continuously flow into sensory processing centers and can overlap or come in at such at increased rate that the demands placed on the working memory are significantly increased and cannot keep up with the rate of information flow (however,

see the section on Top-down processing below) (Grant, Walden, & Seitz, 1998). These limitations in the speed of neural processing may result in the loss of individual sensory information or redundancy in the speech signal, which may affect processes of multimodal integration. Overall, there are a number of physical and psycholinguistic differences across syllables, words, sentences and the many other presentations of speech that listeners encounter, which may change or alter processes of multimodal integration and integration efficiency.

BENEFITS OF AUDITORY-VISUAL INTEGRATION

Visual speech cues enhance and supplement the acoustic characteristics and properties of speech. Summerfield (1987) hypothesized three possible roles for visual cues in improving speech understanding. First, visual cues provide segmental (e.g., consonant and vowel) and suprasegmental (e.g., intonation, stress, rhythmic patterning, etc.) information which is *redundant* with cues provided in the auditory signal, and thus can reinforce auditory information. Second, visual cues provide segmental and suprasegmental information *complementary* to cues provided in the auditory signal (i.e., cues not available in the auditory signal, usually because it is compromised in some way). Third, it is hypothesized that auditory and visual cues share common *spatial and temporal properties* which may help direct auditory attention to speech signals of interest rather than competing speech or background noise. A growing body of research has demonstrated these benefits of auditory-visual speech perception over either speechreading or listening alone (van Wassenhove et al., 2005; Schwartz, Berthommier,

& Savariaux, 2004; Grant & Seitz, 2000; Grant et al., 1998; Grant & Seitz, 1998; Walden, Busacco, & Montgomery, 1993; Blamey, Cowan, Alcantara, Whitfors, & Clark, 1989; Danhauer, Garnet, & Edgerton, 1985; Sumby & Pollack, 1954).

Redundant Information

Visual speech cues provide information that is redundant with the acoustic information found in the speech signal. As described above, movements of the lips and mouth during speech mimic the acoustic speech envelope and frequency transitions in the speech signal (Munhall et. al., 2004; Grant, 2001). Lip and mouth cues also provide information about the place of articulation of speech sounds and lead to the distinction of groups of phonemes (Grant & Seitz, 2000). Some properties of speech, such as those governed by duration, can even be identified by speechreading alone (e.g., certain vowels, stress patterns, etc.). Although visual cues alone can be ambiguous, they provide cues that are redundant with the acoustic properties of speech and enhance the speech signal, improving auditory intelligibility. The additional information can reduce demands placed on the auditory system by speeding up neural processing and decreasing the uncertainty of a perceptual decision (Grant & Seitz, 2000).

As discussed above, studies suggest that synchronous presentation of auditory and visual stimuli can modify the formation of a percept in either modality (van Wassenhove et al., 2005; McGurk & MacDonald, 1976). For example, auditory perception can be enhanced when listeners are provided with visual cues, even when the auditory signal is perfect. Van Wassenhove et al. (2005) showed that visual information greatly enhances the neural processing of auditory information, and that redundant information provided

by visual cues might help the brain to predict auditory utterances, effectively speeding cortical processing.

Complementary Information

As described, visual cues are especially relevant in situations where auditory information is compromised in some way (e.g., loss of hearing, noisy situations, degraded speech) and are used to supplement missing auditory information. Although visual cues do provide information that is redundant with the auditory signal, complete redundancy between the two modalities is uncommon. Voicing information is a relatively robust component of the acoustic properties of speech; however, visual cues supplement acoustic cues by providing the place of articulation information that is easily degraded by background noise or hearing loss. Thus, when speechreading and audition are combined speech perception performance is enhanced. Visual cues allow listeners to decipher consonants and other temporal cues that are lost when the auditory signal has been compromised and supplement missing acoustic information (Grant & Seitz, 2000).

Directed Attention

A growing body of research has provided insight into the benefits of spatial and temporal cues contained in auditory-visual speech (Grant & Seitz, 2002; Grant 2001; Grant & Seitz, 2000; Spence, Ranson, & Driver, 2000). When a listener watches a talker speak, the acoustic and visual cues from the speech signal share common temporal properties, which differ from the acoustic characteristics found in background noise and help direct auditory analysis [attention] to speech signals of interest. Additionally, seeing the location of a talker may help binaural auditory processes to use spatial commonalities to localize and separate speech from competing sound. By providing cues about when

and where to expect auditory information, visual cues can focus the attention of the auditory system on auditory signals of interest and improve a listener's ability to identify speech. It is also suggested that a 1-3 dB improvement in speech detection ability and release from masking occurs with AV speech perception when compared to A perception alone (Grant & Seitz, 2002; Grant 2001; Grant & Seitz, 2000).

AUDITORY-VISUAL INTEGRATION BENEFITS FOR COMPROMISED AUDITORY INPUTS

Summerfield's categorization of the ways that auditory-visual stimuli can lead to enhanced speech perception provides a useful framework for assessing auditory-visual benefit for specific situations in which the auditory signal has been compromised. The benefit that listeners receive from auditory-visual integration during compromised listening situations is of particular interest, because it is in these situations where benefit is maximized. Three such situations are discussed below.

In assessing the effects of compromising the auditory signal it is important to remember the spectral and temporal cues contained in the normal auditory speech signal as well as the information conveyed by these cues. A disruption in the place of articulation, manner, voicing or temporal cues can have a significant impact on speech perception ability, in both auditory and auditory-visual conditions. (Shannon, Zeng, Kamath, Wygonski, Ekelid, 1995).

Hearing Loss

A body of supporting literature suggests that adults with widely differing hearing loss demonstrate improved speech perception performance on tasks of AV speech

recognition when compared to A or V speech alone, regardless of degree, configuration, or duration of hearing loss (Wightman, Kistler, and Brungart, 2006; Schwartz, Berthommier, & Savariaux, 2004; Grant et al., 1998; Grant & Seitz, 1998; Walden, Busacco, & Montgomery, 1993; Blamey, Cowan, Alcantara, Whitfors, & Clark, 1989; Danhauer, Garnet, & Edgerton, 1985). Research has also examined the speech perception performance of adult listeners with hearing loss on various tasks of speech understanding, such as nonsense syllable identification, consonant and vowel identification, and sentence recognition across various conditions of A, V, and AV presentation. Regardless of the speech understanding task, an increase in overall percent correct performance is evident in AV conditions when compared to A or V performance (Wightman et al., 2006; Schwartz et al., 2004; Grant et al., 1998; Grant & Seitz, 1998; Walden et al., 1993; Blamey et al., 1989; Danhauer et al., 1985.).

Helfer (1998) explored whether speaking mode (clear versus conversational speech) was a better auditory stimulus for auditory-visual integration. Clear speech differs from conversational speech in a number of ways. First, speaking rate decreases substantially in clear speech. This decrease is achieved both by inserting pauses between words and by lengthening the durations of individual speech sounds. Second, there are differences between the two speaking modes in the numbers and types of phonological phenomena observed. In conversational speech, vowels are modified or reduced, and word-final stop bursts are often not released. In clear speech, vowels are modified to a lesser extent, and stop bursts, as well as essentially all word-final consonants, are released. Third, the RMS intensities for obstruent sounds, particularly stop consonants, are greater in clear speech than in conversational speech. Finally, changes in the long-

term spectrum are small. Thus, there are clear spectral and temporal differences that exits between clear and conversational speech (Picheny, Durlach, and Braida, 1985).

Results of Helfer's study showed that the speech perception abilities of older adults were approximately thirty percent better when provided with clear AV speech compared to clear A speech and thirty percent better with conversational AV speech compared to conversational A speech regardless of configuration, type, and degree of hearing loss. Results of this study indicate that overall, older adults with hearing loss benefited from AV speech regardless of speaking mode or differences in hearing loss.

A similar study by Grant, Walden, and Seitz (1998) measured syllable and speech recognition ability in a group of older adults with age-related (high-frequency) hearing loss. Recognition of medial consonants in isolated nonsense syllables and of words in sentences in A, V, and AV conditions was evaluated. Participants of this study achieved substantial AV benefit for both sets of materials relative to A recognition performance. Results of this study also suggest auditory-visual communication may address some of the hearing acuity issues experienced by older adults with hearing loss, particularly in consonant identification and sentence recognition.

The vast majority of hearing loss occurs in the higher frequencies where consonant information lies. Consonants are composed of broader spectral content and contain far less energy than vowels. Thus, a high-frequency hearing impairment does not impair the identification of robust vowels sounds in speech, but renders low energy consonant sounds inaudible due to the loss of hearing in those frequency regions. In addition to a loss or decrease in audibility in certain frequency regions, hearing impairment also results in a loss of fine frequency discrimination. Place of articulation

information, which contains cues that allow listeners to distinguish between individual phonemes, is conveyed by subtle changes and fine frequency transitions between more robust vowel sounds. Thus, hearing loss results in a loss of audibility of certain spectral cues and in the ability to distinguish frequency structure and transitions, which impairs a listener's ability to discern individual sounds.

These factors play a big role in a listener's ability to understand speech. As mentioned earlier, consonants convey most of the word information; they are much more important to speech intelligibility than vowels. It is usually possible, for example, to figure out a word if the vowels sounds are removed and only consonants are remaining. However, when consonants are removed speech can sound broken and unintelligible. When place of articulation information is lost, due to hearing loss, listeners are unable to distinguish between individual phonemes and sounds. An additional characteristic of consonants is that they act as breakpoints, separating syllables and words from one another. When these cues are not available, due to hearing loss, individual words and sounds are not clearly defined, sounds run together and speech sounds mumbled. Further complications arise during communication with women and children, since they have higher-pitched voices and are often soft-spoken, and can become completely inaudible to individuals with hearing loss.

As noted by Summerfield, visual cues provide information that is redundant, complementary, or directs attention to the auditory signal, which aids in the perception of compromised auditory inputs and increases the speech perception performance of listeners with hearing loss. For individuals with hearing loss visual cues provide additional vowel, intonation, stress, and rhythmic patterns that are redundant with those

spectral and temporal cues that remain audible to a listener with hearing loss and can enhance saliency of the auditory signal. This information can help listeners to overcome the deleterious effects of hearing loss and discriminate and identify individual sounds and words. Visual cues also provide redundant information conveying manner of articulation, which can enhance those temporal cues that aid in the identification of individual speech sounds. Second, visual cues provide place of articulation information that is complementary to cues available to a listener with hearing loss and can help these listeners to detect and distinguish groups of visemes. This information can also be combined with other available temporal cues to help listeners to distinguish individual phonemes and breakpoints between syllables and words, making speech sounds clear and distinct. There is also a high degree of correlation between lip and mouth area and the second and third formant transitions, which contain mid-to-high frequency spectral information and these cues can be used to further separate visemes and help listeners to discriminate certain phonemes. Visual cues can offer some complementary information about manner of articulation, mostly in timing and duration of sounds. These cues can help listeners to distinguish between certain syllables and words, which may be perceived as “smeared” as a result of hearing loss. Each of these complementary cues can provide information that is missing in the distorted auditory signal and increase speech intelligibility. Third, auditory and visual cues share common spatial and temporal properties; mouth and lip movements mimic the acoustic speech envelope, temporal cues, and frequency transitions in the speech signal and facial movements convey emphatic and phonemic stress, all of which may help direct auditory attention to the crucial components of the acoustic speech signal.

Speech in Background Noise

Research also suggests that visual cues aid in the ability to detect and perceive speech in noise (Munhall et al., 2004; Sumby & Pollack, 1954; Grant & Braida, 1991). One way this benefit is accomplished results from a decrease in the effects of masking when visual cues are available. A recent study by Wightman, Kistler, and Brungart (2006) explored the benefits of auditory-visual integration for children and adults by comparing the relative masking release that could be obtained with the addition of visual cues during speech perception in noise. Results of this study demonstrate a significant improvement in the speech perception abilities of older children (9-16.9 years) and adults in the AV condition when compared to the A condition. Wightman et al., noted that adding visual input to the masked auditory input improved performance at levels comparable to an improvement in the signal-to-noise ratio (SNR) of up to 15 dB. Grant and Braida (1991) also found that listeners achieved significantly higher levels of speech perception in poor signal-to-noise ratios by adding visual cues. Results of this study suggested that adding visual cues to speech in noise may result in an increase in speech perception performance of at least five to ten percent. Furthermore, Munhall et al., (2004) suggest that regardless of the quality of visual information, listeners are able to perceive speech in noise more accurately when visual cues are available. Overall, the literature supports the idea that the addition of visual cues can markedly improve a listener's ability to detect and understand speech in noise (Munhall et al., 2004; Sumby & Pollack, 1954; Grant & Braida, 1991).

Background noise can change in its spectral and temporal properties depending on the type, number, and location of its sources, but typically results in overall masking of

the auditory signal. In the aforementioned studies, speech-spectrum masking noise or multi-talker speech babble were used to compromise auditory information in the speech signal. Background noise is expected to interfere primarily with consonant discrimination and audibility; however, some parts of the speech signal are relatively robust and characteristics such as vowel information, voicing, stress, duration, and other temporal cues may be unaffected by background noise. The loss in frequency specificity and consonant information as a result of background noise leaves speech sounding mumbled, distorted, and broken or inaudible depending on the level and composition of the background noise.

The benefits of the additional visual stimulus are very similar in cases of hearing speech in background noise to those provided to listeners with hearing loss, i.e., restoration of place and manner of articulation information, and directed attention.

Degraded Speech

Pilot studies in our laboratory have offered insight into the benefits offered by visual cues in speech perception performance when the auditory signal has been degraded in some way (Andrews, 2007; Anderson, 2006; ver Hulst, 2006). In these studies auditory stimuli were compromised using different degrees and methods of stimulus degrading. The performance of listeners on congruent speech perception tasks (same word presented auditorily and visually) in degraded A, V, degraded AV, and un-degraded AV conditions were compared.

Results of each of these studies indicate that broad spectral degrading effectively reduces auditory information available in the speech signal. This is supported by decreased percent correct performance in A conditions, regardless of the degree and type

of degrading used to compromise auditory stimuli. Interestingly, although a decrease in speech perception performance due to auditory degrading was noted in each of these experiments, substantial integration was observed in the degraded AV condition, as indicated by an increase in speech perception (percent correct) performance when compared to the degraded A condition. This observation suggests that individuals are able to achieve higher speech perception performance by integrating visual and auditory cues even when there is a loss of information in the auditory signal.

The performance of listeners on tasks of incongruent (e.g., an auditory /gæt/ presented with a visual /bæt/) speech integration was also assessed in each of these studies. Results indicated that degrading the auditory signal has important effects on the multimodal perception process; listeners demonstrate a shift in reliance towards visual cues when the auditory signal is degraded. These results argue that individual dependence on auditory information can be altered by variations in the quality of the auditory input, and that auditory dominance will shift as auditory quality is impaired, increasing the weight of visual cues. In other words, although less weight is given to visual cues in normal speech perception, degrading the auditory signal may help listeners to direct more attention to visual cues when the auditory signal is imperfect in order to achieve optimal speech understanding. Furthermore, the overall amount of auditory-visual integration was similar for normal and degraded AV conditions, suggesting that listeners use all of the sensory information available, regardless of the quality of the information.

Auditory stimuli in each of the studies were degraded using various methods to varying degrees. Auditory degrading consisted of removing or greatly reducing the

spectral structure of the auditory signal while preserving the temporal envelope information in the speech signal. Reduction of the spectral information in the auditory signal resulted in distortion of formant transitions, thus affecting recognition of consonants. The degraded auditory stimuli contained very little information about the place of articulation, relatively little information about the manner of articulation, but by maintaining the temporal envelope retained much of the voicing characteristics of the speech signal. Again, visual cues provided additional information about place of articulation, frequency structure, and breakpoints that can be partially used to overcome the deleterious effects caused by degrading speech.

There is a loss of information in the auditory signal that is unique to each of the situations discussed above, which provides evidence that visual cues can aid in the detection and perception of speech in many complex listening environments and when the auditory signal is compromised in some way.

VARIABILITY IN AUDITORY-VISUAL INTEGRATION AND BENEFIT

Although the research above demonstrates that listeners benefit from combining auditory and visual speech cues, there is a large degree of variability noted across these studies with regard to the amount of integration that individuals achieve during auditory-visual speech perception. Several components will be discussed which may account for the variability in the auditory-visual integration abilities of listeners.

Measures of Integration

Grant and Seitz (1998) hypothesized that the variability seen in auditory-visual integration and benefit may be in part due to the measures of performance being employed in the current research. They examined performance measures for a variety of auditory-visual materials (isolated speech segments, sentences, congruent and incongruent stimuli) to explore whether different measures of auditory-visual integration ability were correlated.

Results of this study showed varying degrees of correlation among different measures of integration and benefit and across different tasks and demands. Auditory-visual integration measures based on the perception of consonants and those based on the perception of key words in sentences differed in the amount of integration and benefit reported. Additionally, measures of auditory-visual benefit for consonant and sentence tests did not reveal significant correlations. When determining the relationship between measures of auditory-visual integration and those of benefit (i.e., integration efficiency, McGurk-type responses, and percent correct performance); little association between these different measures was found. The authors of this study suggest that the variability in the relationship between different measures may be best explained by differences in task demands imposed by different sets of material. Breakdowns observed in the benefit and integration for tests using sentences might suggest that the speed and higher level processing in spoken language recognition may be a separate variable worth examining. However, it is evident from these data that differences in measures of integration and in integration processes for different speech materials may be partly responsible for the variability seen in auditory-visual integration within and across listeners.

Listener Characteristics

Grant and Seitz (1998) also hypothesized that all other things being equal, greater skill at integrating auditory and visual cues, or an individual with a higher integration efficiency, will almost always have better performance on auditory-visual tasks. Highly efficient integrators are assumed to be those individuals who are better at using cues from multiple sources for speech recognition. Differences in listener auditory-visual integration ability may account for some of the variability seen in auditory-visual integration and benefit across listeners.

To explore this hypothesis, Grant and Seitz (1998) examined the differences in auditory-visual integration ability across listeners. Accounting for individual differences in unimodal speech perception ability, they suggested that remaining differences in auditory-visual speech perception are attributable to differing efficiency or ability in the operation of those perceptual processes that integrate auditory and visual speech information. In this study, McGurk-type stimuli produced a greater degree of variability across listeners on all tests when compared to congruent speech stimuli. The authors noted that the presentation of unusual AV materials might have fostered a greater degree of variability than may be observed with more typical speech materials. In particular, the presentation of stimuli that are not consistent with existing linguistic knowledge can lead to confusion. However, when measurements were confined to similar congruent tasks using the same materials, significant effects of subject variability were still observed. Subjects derived substantial benefit from visual speech cues for both consonant and sentence recognition in noise and in the percent correct performance for AV speech materials when compared to A or V speech alone, but there were significant individual

differences in the amount of AV benefit observed. Interestingly, there was no correlation between the benefit that an individual received in identifying syllables in noise and their sentence in noise performance.

Overall, results of this study showed that even when perception tasks, speech materials, and unimodal perception abilities were accounted for, subjects demonstrated a significant amount of variability in the amount of benefit derived from auditory-visual speech. This study supports the notion that individual differences in the perceptual processes that integrate auditory and visual cues, or integration abilities, will in some part explain the relatively large degree of variability in auditory-visual benefit that listeners receive.

Talker Characteristics

There are certain visual characteristics which are known to make an individual a highly intelligible speech talker, for example, facial features, area and shape of lip and mouth opening, eye contact, facial expression, jaw movements, style of speech production, etc. (Jackson, 1988). Stimuli provided by those highly intelligible auditory-visual talkers will almost always yield better performance for listeners on tasks auditory-visual integration and benefit. These talker characteristics may account for some of the variability seen across listeners in auditory-visual speech perception performance.

In one pilot study from our laboratory talker differences were assessed. Subjects were asked to judge degraded A, V, and degraded AV speech syllables, produced by fourteen different talkers. The results of this study showed a significant variability in both auditory intelligibility and auditory-visual integration across talkers.

Interestingly, it was discovered that the AV performance produced by a given talker did not correlate to the A or V performance produced by that same talker.

Overall, results indicated that talker differences play an important role in auditory-visual speech perception. These results suggest that talker characteristics contribute to the variability in the auditory-visual integration seen across listeners; furthermore, those talkers who are highly intelligible auditory or visual talkers may not necessarily possess those characteristics that make them highly intelligible auditory-visual talkers. While this study was a preliminary investigation, results warrant further exploration of talker characteristics and the role that they play in the auditory-visual integration process.

Signal Characteristics

An alternative explanation for the variability observed in the auditory-visual integration abilities of listeners may be the characteristics of the auditory signal. The natural speech signal is redundant in that it contains far more information that minimally necessary for successful speech perception. Shannon, Zeng, Wygoski, Kamath, and Ekelid (1995) examined how the amount of information available in the auditory signal affected speech perception. Listeners were presented with degraded speech stimuli containing different amounts of information and their speech perception abilities were measured. The stimuli were degraded using different numbers of bandpass filtered speech: 1-channel, 2-channels, 3-channels, and 4-channels. Speech stimuli were further removed of temporal fine structure that was replaced by broad band noise which was modulated by the original temporal envelope. Results of this study revealed the identification of consonants, vowels, and words in simple sentences improved markedly

as the number of bands increased; high speech recognition performance was obtained with only three bands of modulated noise. Thus, the presentation of a dynamic temporal pattern in only a few broad spectral regions is sufficient for the recognition of speech. Results of this study demonstrate the redundancy of the natural speech signal.

When the auditory signal is compromised in some way, it changes the amount of information and quality of the speech signal that a listener receives. The changes in the speech signal may affect a listener's ability to extract auditory information or integrate auditory and visual cues. It is likely that differences in the auditory signal contribute to the variability seen in auditory-visual integration processes.

Pilot studies from our laboratory have explored auditory-visual integration in situations where the auditory signal is compromised in some way (Andrews, 2007; Tamosiunas, 2007; Anderson, 2006; ver Hulst, 2006; Yeary, 2001). These studies assessed whether listeners are able to integrate auditory-visual cues when information has been removed from the auditory signal, and have examined the reliance on the auditory signal during speech perception when it is compromised.

As previously mentioned, these studies suggest that regardless of the method of stimulus degrading, increases in speech perception performance were observed in AV conditions over A conditions, indicating that listeners do in fact integrate auditory and visual information when the auditory signal is compromised and that listeners benefit from this integration. Listeners in these studies also demonstrated a shift in reliance towards visual cues when the auditory signal was degraded, as demonstrated by the results of discrepant McGurk-type tasks. These results support the idea that individual dependence on auditory information can be altered by variations in the quality of the

auditory input, and that auditory dominance will shift as auditory quality is impaired, increasing the weight of visual cues.

Ross, Saint-Amour, Leavitt, Javitt & Foxe (2006) presented evidence that auditory-visual integration varies depending on the quality and quantity of auditory information available in the speech signal. Listeners in this study were presented with speech syllables in background noise at varying levels of signal-to-noise ratio (SNR) and asked to identify the speech syllable stimuli, in both A and AV conditions. Results of this study indicated that optimal auditory-visual integration was achieved at intermediate signal-to-noise ratios. The authors suggested that a certain amount of auditory information is required before auditory-visual integration occurs and speech perception can be enhanced by visual cues. However, it was also suggested that there is a certain point where visual cues cannot further enhance speech perception. It is likely that at increased SNRs the additional visual cues are adding information that is more redundant to the highly intelligible auditory cues rather than providing any new information. Although listeners received benefit from the addition of visual speech while it was available, in extreme (positive or negative) SNRs, one modality was clearly dominant over the other and listeners relied less heavily on the integration of cues across modalities to understand speech. Although some models of auditory-visual integration suggest that dominance of one modality will shift along a continuum depending on the quality and strength of the two inputs, Ross et al. suggest that information from the less-dominant modality is discarded when enough information is contained in the dominant signal. Results of this study provide support that auditory signal characteristics may be in part

responsible for the differences in the amount of auditory-visual integration seen across listeners.

Another recent study by Grant, Tuft, and Greenburg (2007) evaluated the speech perception abilities of normal-hearing and hearing-impaired listeners to determine the degree to which information available in the auditory signal may be a factor in speech perception performance. Nonsense speech syllables consisting of eighteen medial consonants surrounded by the vowel /ɑ/ were degraded to varying degrees. The stimuli were degraded using 1, 2, 3, and 4 bandpass filter channels. The temporal fine structure was then removed from the speech syllable stimuli and was replaced with narrowband noise, modulated by the temporal envelope retained in the filtered speech syllable. Performance was examined in three different conditions: A, V, and A+V. Results of this study yielded two very interesting findings. First, A recognition performance for the hearing-impaired listeners was worse than that of the normal-hearing listeners for all levels of stimulus degrading, which is to be expected given their residual hearing deficits. Interestingly, the data showed that V and AV perception performance was comparable across subject groups for all levels of stimulus degrading, suggesting that most of the hearing deficit was overcome when visual cues were combined with even limited auditory information. Auditory degrading affected speech perception in the A condition for hearing-impaired listeners, but did not affect their performance in the AV condition when compared to normal-hearing listeners and results showed that both normal-hearing and hearing-impaired listeners integrated information across the auditory and visual modalities, independent of differences in auditory capabilities. However, results revealed that hearing impaired listeners received a greater amount of benefit from auditory-visual

integration. The second finding worth noticing is that both normal-hearing and hearing-impaired listeners achieved the same amount of auditory-visual integration across all levels of stimulus degrading.

Interestingly, results of this study yielded contradicting and similar findings compared to those noted by Ross et al. (2006). In this study, the hearing-impaired listeners consistently achieved a greater amount of auditory-visual integration when compared to the normal-hearing listeners across all levels of degrading. When greater amounts of information were removed from the auditory signal listeners received a greater amount of benefit from auditory-visual integration; the additional loss of information from the hearing loss resulted in greater amount of information being "removed" for this group of listeners. As opposed to the findings of Ross et al. the results of this study imply that optimal auditory-integration is achieved when there is less information available in the auditory signal. A second point worth discussing is in examining the amount of auditory information for each of the groups separately findings revealed that the amount of auditory-visual integration remained consistent across the varying levels of stimulus degrading. As suggested by Ross et al., a certain amount of auditory information must be removed before auditory-visual integration occurs and speech perception can be enhanced by visual cues. It is possible that the amount of information available in the auditory signal across all levels of degrading was not sufficiently different to produce a change in auditory-visual integration ability, but only after the compounding hearing loss were there significant differences in the amount of auditory information available offering increased benefit by the new information afforded by the visual cues.

Results of these two studies yielded some similar and conflicting findings regarding the auditory signal and its role in auditory-visual speech integration. Although both studies suggest that the auditory signal does in fact play a role in this process it has not yet been determined what amount of information is necessary for listeners to achieve optimal benefit from auditory-visual integration. Further investigation of the auditory signal and its role in auditory-visual integration is warranted.

ROLE OF THE AUDITORY SIGNAL IN AUDITORY-VISUAL INTEGRATION

Clearly, there is a substantial degree of variability in auditory-visual integration abilities across listeners. It has been suggested that differences in the information contained in the auditory signal might be partly responsible for this variability. The primary purpose of the present study was to determine if the amount of information available in the auditory signal has an affect on listeners' abilities to integrate auditory and visual cues. A secondary purpose of the present study was to determine how changes in the amount of information available in the auditory signal affects the auditory-visual integration process as well as to provide the footwork for further examining the role of the auditory signal as a potential source for the variability in the benefit listeners receive from auditory-visual integration.

The present study examined listeners' auditory-visual integration abilities by evaluating and comparing their speech perception performance in auditory-only (A), visual-only (V), congruent auditory + visual (AV), and incongruent auditory + visual (AV) conditions using auditory stimuli that were degraded to varying degrees. The

stimuli were degraded using 2, 4, 6, and 8 bandpass filter channels. The temporal fine structure was then removed from the speech syllable stimuli and was replaced with narrowband noise, modulated by the temporal envelope retained in the filtered speech waveform. AV benefit was compared across all levels of auditory signal degrading to examine how the differing amounts of information available in the auditory signal impacted the benefit that listeners achieved from integration.

The present study explored the possibility that removing information from the auditory speech signal might change multimodal integration. Additionally, we examined how the integration process is affected when the amount of information in the auditory signal changes and attempted to determine the amount of information necessary for optimal auditory-visual integration. It was anticipated that auditory signal degrading would impact performance in the A and AV conditions similarly to findings noted in the current literature in multimodal speech perception. It was also expected that a certain amount of auditory information would be necessary for auditory-visual integration to occur and that a certain amount of information must be missing from the auditory signal before visual cues would offer additional information and speech perception could be enhanced by these cues. We also expected that as greater amounts of information were removed from the auditory signal listeners would receive a greater amount of benefit from auditory-visual integration and that optimal auditory-integration would be achieved when less information was available in the auditory signal.

CHAPTER 3

METHOD

PARTICIPANTS

Four male and sixteen female participants ranging from nineteen to twenty-three years of age participated as listeners in the present study. Two male and three female participants ranging from twenty to twenty-three years of age participated as talkers in the present study. All talkers and listeners were native speakers of Midwestern American English, per participant report. All observers were screened for normal hearing bilaterally using a criterion of 20 dB HL at 500 to 8000 Hz and reported normal or corrected-to-normal vision. Eight of the observers were undergraduate students in Speech and Hearing Science and the remaining twelve observers had varying disciplines of undergraduate study. Four of the participants majoring in Speech and Hearing Science reported some knowledge of the McGurk effect.

STIMULI

Stimulus Selection

The stimuli in the present study were chosen to satisfy the following criteria:

1. The stimuli differed only in initial consonant (minimal pairs).

2. All stimuli included the same vowel, /æ/, since it does not include lip rounding or extension.

3. The stimulus set included a good representation of each category of articulation: place (bilabial, veolar), manner (stop, fricative, nasal), and voicing (voiced, unvoiced).

4. Stimuli were known to elicit McGurk responses.

Stimulus Set

The stimulus set for this experiment consisted of eight CVC syllables. The stimulus set was presented as both congruent (same syllable in both auditory and visual modalities) and discrepant-syllable (different syllables in auditory and visual modalities, known to elicit McGurk-type responses) stimuli.

Congruent Syllable Stimuli

/bæt/, /pæt/, /mæt/, /sæt/, /zæt/, /tæt/, /gæt/, /cæt/

Discrepant Syllable Stimulus Pairs

Presented as visual-auditory

/cæt-pæt/, /pæt-cæt/, /bæt-gæt/, /gæt-bæt/

STIMULUS PRODUCTION

Audio Signal Degrading

Auditory and visual syllables were recorded (via computer and digital video camera) until each talker provided five usable productions of each syllable for auditory degradation and visual speech images. The original speech syllables were recorded

through a microphone fed directly into a computer, which allowed for the auditory files to be stored in .wav format. These auditory files were then degraded in a manner similar to that used by Shannon et al. (1998), using a subroutine created by Bertrand Delgutte in MATLAB 5.3. The subroutine ("Chimeras") degraded the speech syllables by creating a waveform composed of a broadband noise fine structure that was modulated by the temporal envelope of filtered speech syllables. The stimuli were degraded using 2, 4, 6, and 8 bandpass filter channels. The temporal fine structure was then removed from the speech syllable stimuli and was replaced with narrowband noise, modulated by the temporal envelope retained in the filtered speech waveform. Each degraded speech signal was filtered into a specified number of bandpass channels. The bandwidths of the bandpass filters were selected within the program to simulate non-linear basilar membrane organization. The cutoff frequencies for each of the bandpass filtered channels are as follows:

2 Channel 80 Hz - 1,877 Hz - 19.2kHz.

4 Channel 80 Hz - 518 Hz - 1,877 Hz - 6,097 Hz - 19.2 kHz.

6 Channel 80 Hz - 315 Hz - 814 Hz - 1,877 Hz - 4,139 Hz - 8,953 Hz -19.2 kHz.

8 Channel 80 Hz - 238 Hz - 518 Hz - 1,010 Hz - 1,877 Hz - 3,404 Hz - 6,097 Hz - 10,840 Hz -19.2 kHz.

Digital Video Editing

Visual stimuli for the study were obtained by recording each of the talkers repeating the speech syllables chosen for the stimulus set with a digital video camera. Digital video recordings were then downloaded and edited using a computer software program, Video Explosion Deluxe. Within this program the auditory signal from the

digital video recording was removed and discarded. The degraded auditory stimuli created within the "chimeras" subroutine were then dubbed onto the visual speech image. This editing software made it possible to create audio-visual stimuli that featured a normal visual representation with a degraded auditory signal and also permitted the formation of discrepant (McGurk-type) stimuli. For the present study the visual stimuli produced by a talker were paired only with the degraded auditory stimuli produced by that same talker. The program Sonic MY DVD was used to burn stimulus lists created in Video Explosion Deluxe to DVD.

Stimulus Lists

Three randomized stimulus lists were created for each talker in each of the four conditions, resulting in a total of sixty different stimulus lists for stimulus presentation. Randomized stimulus lists were created in order to reduce the possibility of effects that can occur from order of stimulus presentation. In the A and V conditions each stimulus list included sixty speech syllable stimuli. In the A + V conditions each stimulus list included thirty congruent speech syllable stimuli and thirty discrepant speech syllable stimuli. All syllables were presented in isolation without a carrier phrase.

INTERFACES FOR STIMULUS PRESENTATION

Visual Presentation

Visual stimuli were presented via a 20" television monitor connected to a DVD player.

Auditory Presentation

Auditory stimuli were presented from the headphone output of the television monitor to TDH 39-Audiologic headphones at approximately 50 dB.

METHODS FOR MEASURING AUDITORY-VISUAL INTEGRATION

Percent Correct Performance

Percent correct performance in the A and AV conditions was measured in order to quantify the increase in performance (amount of auditory-visual integration) in the AV condition (AV-A).

McGurk Susceptibility

Another measure of integration is McGurk susceptibility. Discrepant stimuli which are known to elicit McGurk-type integration were presented and listener responses were analyzed for the rate of fusion and combination McGurk-type responses across all level of stimulus degradation. A fusion response occurs when the listener blends the place of articulation information from the two syllables and perceives an entirely new place of articulation somewhere between the two, for example an auditory /bæt/ presented with a visual /gæt/ would result in the listener perceiving the syllable /dæt/. A combination response occurs when the listener combines the place of articulation information from the two original stimuli together, for example an auditory /bæt/ presented with a visual /gæt/ would result in the listener perceiving the syllable /bgæt/. The role that the auditory signal plays in auditory-visual integration was explored by varying the levels of stimulus degradation and analyzing changes in AV integration.

PROCEDURE

Presentation Condition

All participants observed each of the five talkers in the following three conditions: visual only (V), degraded auditory only (A), and degraded auditory + visual (AV).

Presentation Level

Auditory stimuli in the A and A + V conditions were presented under four levels of stimulus degradation: 2, 4, 6, and 8 channels of bandpass filtered speech.

Testing Set-Up

Testing for this study was conducted in a quiet environment with fluorescent lighting. Testing was conducted in a single walled sound-attenuating booth, with the door sealed to reduce ambient noise. Participants were seated in a chair positioned along the back wall of the sound booth facing a glass window in the booth through which they were able to view the television monitor placed outside. While seated, the participants were roughly four feet from the television monitor. The window shade was pulled down for auditory alone conditions and raised for all other conditions. Participants wore headphones in all conditions that utilized an auditory stimulus. Subject responses were transmitted through an intercom system in the booth to an examiner in the control room.

Presentation Task

Two different listening tasks were assessed during AV presentation. “Same trials” included congruent speech stimuli featuring the same degraded auditory syllable and visual syllable paired together for presentation and “different trials” featured discrepant speech stimuli consisting of different degraded auditory and visual syllables paired

together for presentation using combinations of syllables which are known to elicit McGurk type responses.

Testing Procedure

Each participant was provided with written and verbal instructions for testing. Additional questions that were judged to provide further information about the study or add bias were prohibited. The instructions informed the listeners that they would be presented with speech syllables spoken by a number of different talkers and would be asked to judge what syllable they perceived during each presentation. The listeners were informed that they would be presented with degraded speech syllables in three randomized presentation conditions; A, V, and AV. The participants were instructed that each of the stimuli were minimal pairs ending "at" and that any beginning consonant or combination of consonants was a valid response. The participants were also instructed that the speech syllables did not have to form a known word and could be a nonsense syllable or consist of a combination of syllables found in languages other than English. The participants were told to respond to each of the sixty stimuli on each DVD by repeating the syllable that was perceived.

Each listener was tested over approximately twelve hours, in multiple sessions that lasted between one and two hours each. Frequent rest periods were provided to minimize fatigue. One week of training was provided to ensure that participants understood the task and that best performance was judged. All participants were tested in each of the three presentation conditions at each of the four levels of auditory degradation for each of the five talkers. The presentation order of talker, condition, and level was randomized for all participants. No feedback was provided during training or testing.

DATA ANALYSIS

Percent correct and percent McGurk responses were arcsine transformed for parametric analysis. Data analysis compared the speech percent correct performance for congruent speech stimuli in A, V, and AV conditions, as a function of different levels of auditory degradation (2, 4, 6, and 8 channels) to analyze performance differences. A 2-factor ANOVA using a repeated measures design was used to measure main effects of stimulus condition and number of channels, as well as interaction effects. Specific means comparisons were performed to identify sources of significant differences. Finally, the proportion of variance accounted for was computed to determine the strength of any significant effects. Additionally, a descriptive comparison of the percentage of McGurk responses under the different levels of auditory degradation was performed.

CHAPTER 4

RESULTS AND DISCUSSION

The results for two different types of stimuli were analyzed. First, performance was evaluated for congruent stimuli (same auditory and visual stimulus) and percent correct performance was measured across all conditions (degraded auditory-only, visual-only, degraded auditory + visual) at all level of degradation (2, 4, 6, and 8 channels). The degree to which the auditory + visual performance improved over the auditory-only or visual-only performance served as a measure of integration and auditory-visual benefit.

Second, discrepant stimuli (different auditory and visual stimulus) were assessed. These responses were not recorded for percent correct, because there is no "correct" response for the differing stimuli. These responses were recorded into three categories: auditory (the response was the same as the auditory stimulus), visual (the response was the same as the visual stimulus), and other (the response differed from both the auditory and visual stimuli) and analyzed for evidence of integration of the differing stimuli.

PERCENT CORRECT PERFORMANCE

Figure 1 shows the percent correct identification for visual-only (V), auditory-only (A), and congruent auditory + visual (AV) conditions across all levels of auditory signal degrading (2-, 4-, 6-, and 8-channels). Results are averaged across talkers and listeners.

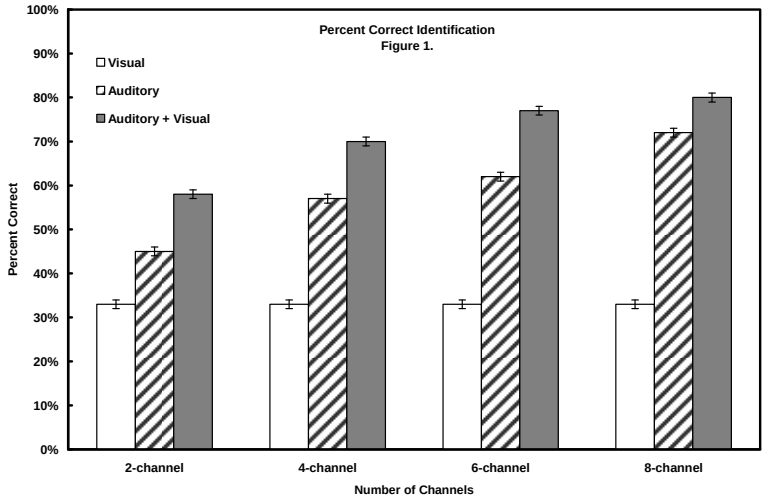

Figure 1. Percent correct identification for 2-, 4-, 6- and 8-channel stimuli in the visual-only (V), auditory-only (A), and auditory + visual (AV) conditions.

Figure 1 reveals several interesting findings. The first point worth noting is that V performance is consistent across all levels of degradation, which is to be expected since the varying factor among the different presentation levels is only in the available amount of auditory information and should not affect the V performance. The average V performance is consistent with numerous pilot studies conducted in our lab and speaks to the validity of the present study (Anderson, 2007; Andrews, 2007; ver Hulst, 2006).

The second finding worth noticing is that A and AV performance systematically increases with the number of output channels. Statistical analysis revealed a significant main effect of number of channels, $F(3, 294) = 123.226$, $p<.001$. Follow-up pairwise comparisons showed significant differences among all pairs of channels, except between 4- and 6-channels.

In the A condition speech recognition improved 12% from 2-channels to 4-channels, 5% from 4- to 6-channels, and 10% from 6- to 8-channels. In the AV condition speech recognition improved 12% from 2-channels to 4-channels, 7% t from 4- to 6-channels, and 3% from 6- to 8-channels. This finding implies that in both that A and AV conditions listeners are able to take advantage of the additional auditory information available as the number of output channels increases and use this information to improve speech perception.

This finding also suggests that auditory stimuli reduced to a 2-channel output contain substantially less auditory information than 4-, 6-, or 8- output channels, resulting is significantly poorer speech perception performance. Additionally, it could be implied that 4- and 6- output channels contain similar amounts of auditory information, as there was no statistical difference between performances at these levels. Finally, it can be inferred that 8-channels contain far more auditory information than 2-, 4-, and 6-channels, resulting in significantly better listener performance for this stimulus.

Auditory signal degrading clearly decreases speech perception performance in both the A and AV conditions when compared to the performance of listeners in previous studies in our laboratory in which listeners achieved 100% correct performance on similar A and AV tasks when the auditory signal was not degraded (Anderson, 2006; ver

Hulst 2006). Nonetheless, listeners remain able to integrate auditory and visual cues across the varying levels of auditory signal degrading. These findings support the notion that a loss of information in the auditory signal does not completely inhibit listeners from integrating auditory and visual cues.

Results also revealed a significant main effect of presentation condition, $F(2, 196) = 665.86$, $p<.001$. Pairwise comparisons showed significant differences among all three of the presentation conditions. Figure 1 shows that listeners exhibit the poorest speech perception performance in the V condition and the best performance in the AV condition across all degradation levels. These results suggest that the visual stimulus contains substantially less usable information for discerning speech, when compared to an auditory or auditory-visual stimulus resulting in significantly poorer speech perception performance in this condition. Listeners are able to benefit from the presentation of auditory-visual information, given that performance in the AV condition exceeded the performance in both the V and A conditions across all levels of auditory degradation. Finally, a significant interaction of number of channels and presentation condition, $F(6, 588) = 46.69$, $p<.001$, was found, reflecting the flat V performance across degradation levels.

The results in Figure 1 were anticipated and similar to findings noted in the current literature by Grant et al. (2007) and Ross et al. (2006), which indicate that auditory signal degrading negatively impacts speech perception performance in the A and AV conditions.

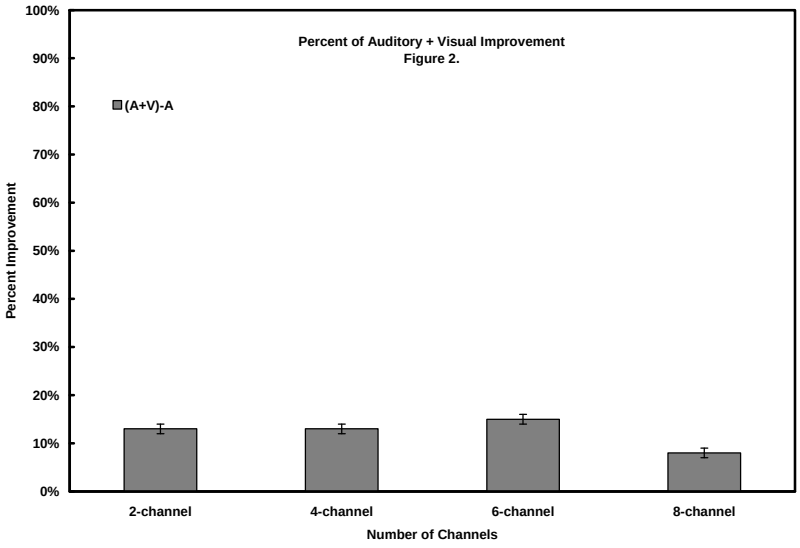

Figure 2. Percent auditory + visual improvement for 2-, 4-, 6- and 8-channel stimuli

Figure 2 displays the percent improvement across talkers and listeners in the AV condition over the performance in the A condition by number of channels, which has been used as a measure of auditory-visual integration. For the 2-channel as well as the 4-channel stimuli, the AV condition showed a 13% improvement over the A condition. For the 6-channel stimuli there is a 15% improvement and there is 8% improvement at the 8-channel stimuli. This increase in performance is relatively similar across all levels of degradation with a slight decrease at the 8-channel level. The minimal decrease at this level suggests that the visual stimulus may not provide as much additional information, because the auditory signal already contains a sufficient amount of information and is highly intelligible. Results revealed significant differences in the amount of auditory-visual integration, $F(3, 297) = 5.339$, $p = .041$, however, the strength of this significance was weak and pairwise comparisons revealed that only the 8-channel stimuli were statistically different from the remaining stimuli.

Overall, these results show that the addition of visual cues provides new information to the speech signal that is not available in the A condition and that listeners benefit from this additional information as seen by an improvement in their speech perception performance. It was expected that a certain amount of auditory information would be necessary for auditory-visual integration to occur and that a certain amount of information must be missing from the auditory signal before visual cues would offer additional information and speech perception could be enhanced by these cues.

However, despite the significant difference for the 8-channel stimuli, auditory-visual integration remained relatively consistent across all levels of stimulus degrading no matter how the amount of information available in the degraded auditory signal was altered, which was contrary to the expected outcomes of the present study. This finding was surprising, due to our expectation that as greater amounts of information were removed from the auditory signal listeners would receive a greater amount of benefit from auditory-visual integration and that optimal auditory-integration would be achieved when less information was available in the auditory signal.

These findings support to some degree the findings noted by Grant et al. (2007) and Ross et al. (2006). The results of this study do suggest that there is some point at which a loss of information in the auditory signal results in decreased AV performance, but auditory-visual integration remains unchanged. Clearly, the loss of information in the auditory signal does impact AV performance, because listeners are unable to achieve perfect speech perception performance at any level of degradation in the AV condition. It is possible that type of auditory signal degrading employed in the present study did not remove substantially different amounts of information from the 2, 4, 6 and 8 channel

stimuli to produce differences in auditory-visual integration across the levels of degradation, a finding that is similar to those noted by Grant et al. (2007). However, it is also possible the information that was removed from the auditory signal by degrading was not provided by the visual cues and thus, relatively little new information was added and speech perception was not further enhanced.

A secondary set of analyses was conducted in order to determine whether variability across talkers might have obscured a possible impact of reduced auditory information on integration performance. Figures 3 and 4 show the percent correct identification in the A and AV conditions by talker, respectively. To assess differences, separate ANOVAs were performed in each condition to evaluate main effects of talker and level of auditory stimulus degradation.

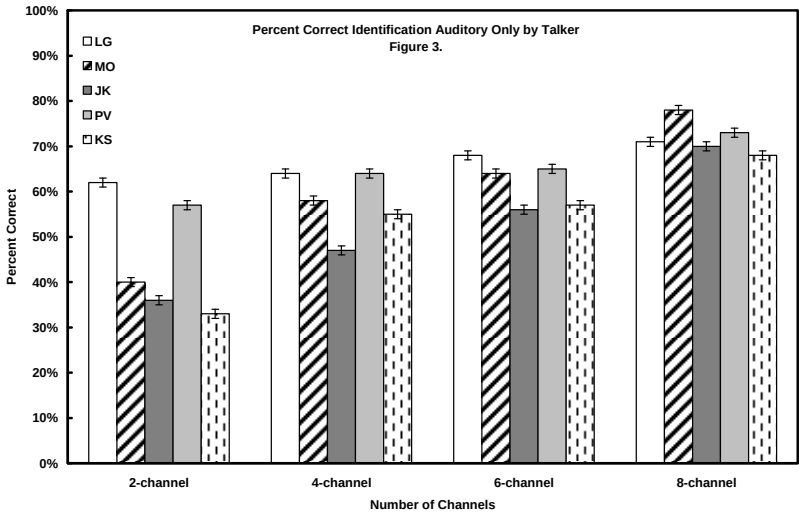

Figure 3. Percent correct identification auditory-only by talker for 2-, 4-, 6- and 8-channel stimuli

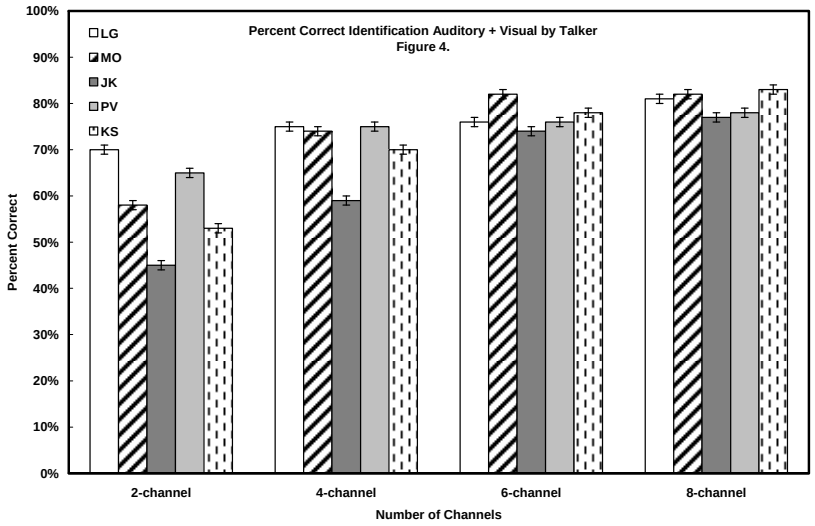

Figure 4. Percent correct identification auditory + visual by talker for 2-, 4-, 6- and 8-channel stimuli

These figures show a great deal of variability across talkers for the 2-channel stimuli. Talkers LG and PV are more intelligible than the other three talkers. MO, JK, and KS are each slightly more intelligible than each other, respectively. There was a significant main effect of talker, $F(4,76) = 33.31$, $p<.001$. Pairwise comparisons showed that LG and PV performed significantly differently from the remaining talkers. There was also a significant main effect of level of degradation, $F(3, 54) = 65.86$, $p<.001$, as well as an interaction effect of talker and level of degradation, $F(8, 152) = 11.243$, $p<.001$.

For the 4-channel stimuli, there was also a great deal of variability, with talkers LG and PV slightly more intelligible than the other talkers; however, MO was more intelligible than talkers JK and KS. In the 4-channel analysis, there was again significant main effect of talker, $F(4,76) = 9.47$, $p<.001$. Pairwise comparison showed that talker JK performed significantly worse than the remaining talkers. There was also a significant

main effect of level of degradation, $F(3, 54) = 62.85$, $p<.001$, as well as There was also a significant interaction effect of talker and condition, $F(8, 152) = 3.913$, $p<.001$.

For the 6- channel stimuli, the gap between percent correct performances across talkers began to close, and for the 8-channel stimuli variability is no longer significant. 8-channel results show a high level of intelligibility across all talkers. For the 6-channel analysis, there was a significant effect of talker, $F(4, 76) = 4.8$, $p=.003$. For the 6-channel stimuli, pairwise comparisons revealed that LG, MO, and PV were significantly different from JK and KS. There was also a significant main effect of level of degradation, $F(3, 54) = 54.26$, $p<.001$, $r^2 = .35$ as well as significant interaction effects of talker and level of degradation, $F(8, 152) = 2.410$, $p=.023$. In the 8-channel analysis there were no differences or main effects across talkers.

These results suggest that when a limited amount of information is available in the auditory signal there is a great deal of variability in listener performance across talkers. As the information in the auditory signal increases, performance variability decreases. It is also possible that the variability seen across talkers for the 2-, 4- and 6-channel stimuli may have contributed to the lack of significant differences in auditory-visual integration as a function of the auditory signal.

Another interesting finding worth noting is that for the 2-channel stimuli, talkers MO and KS produced the greatest amount of auditory-visual integration. For the 4-channel stimuli, talkers MO, KS, and JK produced the most integration. These results suggest that the worst talkers in the A condition might produce the most benefit in the AV condition.

McGurk-Type Integration

Analyses of McGurk-type integration can offer the ability to compare the auditory-visual integration occurring with un-degraded and degraded stimuli, permitting exploration of the role the auditory signal plays in auditory-visual integration. The results above suggest that removal of information from the auditory signal does not impact the amount of auditory-visual integration observed. However, differences in integration between un-degraded and degraded stimuli may yield a different view of the role of the auditory signal and this process.

Figure 5 shows the percent of visual, auditory and other (a response that is not the visual or auditory stimulus) responses for the discrepant stimuli by the number of channels.

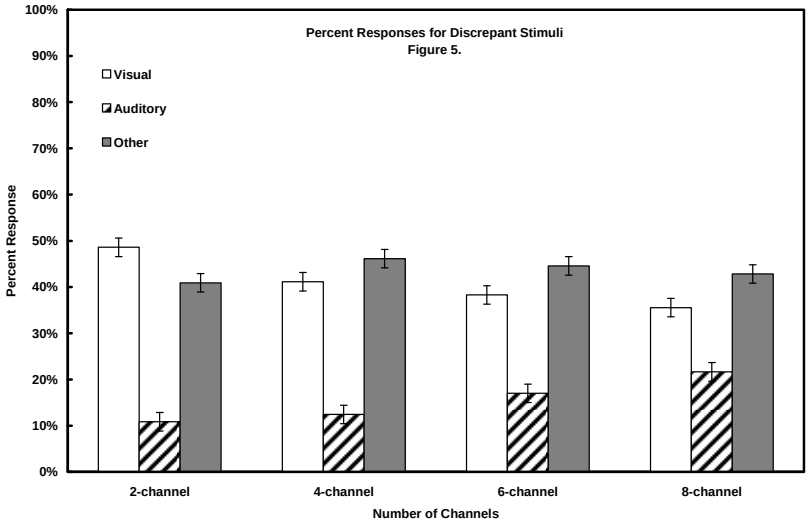

Figure 5. Percent modality-based responses for discrepant stimuli for 2-, 4-, 6- and 8-channel stimuli

The visual response was the highest percentage of responses at the 2-channel level, but showed a slight systematic decrease (49% for 2-channel, 41% for 4-channel,

38% for 6-channel, and 36% for 8-channel) as the number of auditory channels increases. Visual responses were greater than auditory responses for all auditory stimulus configurations. Thus, despite the relative ambiguity of the visual signal, subjects were more likely to rely on visual cues for information rather than the degraded auditory signal. This finding appears to reflect a change in subject behavior based on the relative strength of the auditory and visual inputs, such that listeners in the present study have a shift in the modality relied upon for speech perception as a function of auditory input strength. These results suggest that filtering the auditory signal by 2-, 4-, 6-, and 8-channels removes a sufficient amount of information from the auditory signal to cause listeners to shift their reliance to the visual modality during speech perception.

Auditory responses were consistently the lowest percentage of responses across all channels (11% for 2-channel, 12% for 4-channel, 17% for 6-channel, and 22% for 8-channel); however, there is a slight increase in the auditory response rate as the number of channels increases. This finding might suggest that as additional information is available in the auditory signal listeners rely more on the auditory input during speech perception and again indicates that modality salience plays a role in perceptual decision making. However, the minimal number of auditory responses and lack of significant differences across channels is surprising given the increase in available auditory information as the number of channels increases, as shown by the high levels of percent correct performance in identification in the 6- and 8- channels in the A and AV conditions (see Figure 1).

The “other” response rates were fairly consistent across all stimulus configurations (41% for 2-channel, 46% for 4-channel, 45% for 6-channel, and 43% for

8-channel) and make up the highest percentage of total responses. These results indicate that listeners were affected by the loss of information in the auditory signal and behaved in one of two ways: 1) because the quality of the auditory information was significantly reduced listeners shifted their reliance on auditory cues for speech perception and relied more heavily on visual cues or integrated cues. Since the visual speech signal is inherently ambiguous a combination of the auditory and visual cues might have provided the most information for listeners for speech perception resulting in a shift of reliance towards integration of these two modalities, 2) or discrepant stimuli of rather ambiguous signals caused listeners to guess what they heard making a non-integrated response; both resulting in an "other" response.

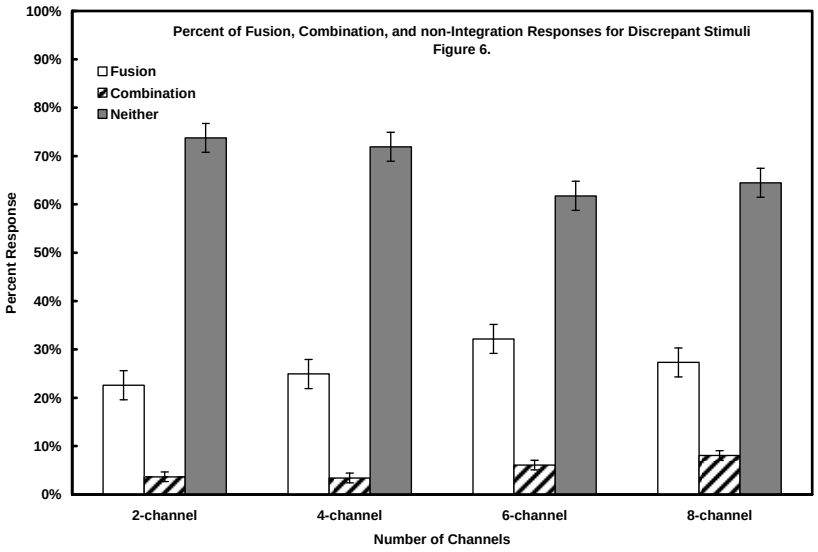

Figure 6. Percent fusion, combination, and non-integrated responses for discrepant stimuli for 2-, 4-, 6- and 8-channel stimuli

Figure 6 shows a breakdown of the "other" responses in Figure 5 to assess the amount of auditory-visual integration that occurred in the discrepant stimulus trials. The

"other" category was analyzed to determine if the individual responses could be classified as fusion or combination McGurk-type integration of the auditory and visual stimuli or if they were non-integrated responses. In the present study, a McGurk-type combination was defined as a response that combined the initial consonant phonemes from the auditory and visual stimuli to form a consonant cluster of those phonemes; for example, visual /bæt/ presented with auditory /gæt/ would result in the response-/bgæt/, a combination of the two stimuli. A McGurk-type fusion was defined as a response that blended the initial consonant phonemes from the auditory and visual stimuli resulting in a shift in the place of articulation between the two original place locations forming a new initial consonant phoneme; for example, visual /gat/ presented with auditory /bæat/ would result in the response /dæt/, a fusion of the two stimuli. Any response that was not classified as a fusion or combination response was considered to be a non-integrated response.

The lowest percentage of "other" responses was the McGurk-type combination responses. This finding is commensurate with previous findings in auditory-visual integration; combination response rate is assumed to be low due to the fact that these consonant clusters are not found in Standard American English and thus, listeners are less likely to produce them as a response.

Auditory-visual fusion responses, although elevated when compared to combination responses, were low across all channels. This finding was unexpected in comparison to previous studies in this laboratory with the same stimulus set using undegraded auditory stimuli in which fusion integration rates using were near 50%-60% of all responses. In the present study, fusion responses consisted of 23% of total

responses at the 2-channel, 25% at the 4-channel, 32% at the 6-channel, and 27% at the 8-channel level, a substantial decrease in the number of fusion responses across all channels. These results suggest that removing information from the auditory signal can negatively impact the auditory-visual integration of discrepant stimuli.

McGurk and MacDonald (1976) found that adult listeners integrate discrepant stimuli about 90-92% of the time. The much lower levels of integration observed in the present study suggest that removing information from the auditory signal may impact the auditory-visual integration of these stimuli. This finding might also indicate that a loss of information in the auditory signal may affect the integration of different types of auditory stimuli (congruent vs. discrepant) or in different situations in different ways. Such a notion is plausible given anecdotal observations that McGurk-type stimuli are inherently somewhat ambiguous.

One concern with regard to the validity of the "other" responses was the surprisingly large number of /hæt/ responses, which were produced by all subjects across all stimuli. The phoneme /h/ was not classified as a fusion or combination response, because the location of production (glottal) does constitute a blend in the bilabial and the velar places of articulation found in the original McGurk stimuli. It is possible that noise added by the individual transducers (DVD player, amplifier, TV and headphones) may have contributed to the high rate of /hæt/ responses. In order to investigate this, five listeners were re-tested with stimuli presented on a computer with direct audio output. Results obtained from the computer trials were inconclusive regarding the transducer effects in the present study; even though all but one subject had fewer /hæt/ responses for the computer trials differences in /hæt/ responses between the test booth and the

computer trails varied greatly among subjects. Both AS and MG had one less /hæt/ response for the computer trial, whereas MT and KB had a substantial decrease in /hæt/ responses. NO produced five additional /hæt/ responses for the computer trials that were not produced during the test booth trials. Also, for AV presentations on the computer, the /hæt/ responses were all produced for McGurk stimuli. This is an interesting discovery that may imply integration as opposed to noise interference. Further study is needed to determine whether transducer affected the perception of auditory stimuli in the present study.

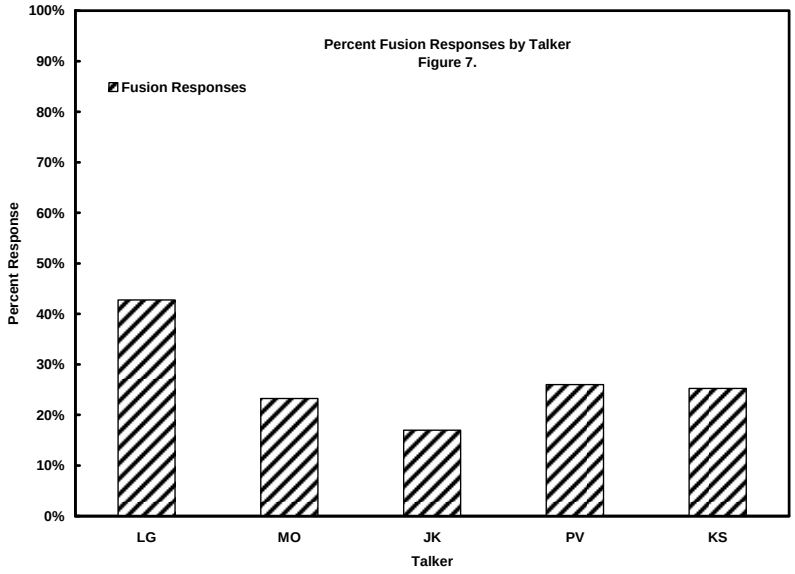

Figure 7. Percent fusion responses by talker

Figure 7 investigates fusion responses across talkers. This figure reveals noticeable variability in the fusion response patterns across talkers, specifically, the differences between talker LG, who has a higher fusion rate when compared to the other

talkers, and JK, who has a very low rate. Results suggest that talker characteristics also affect fusion integration of these stimuli.

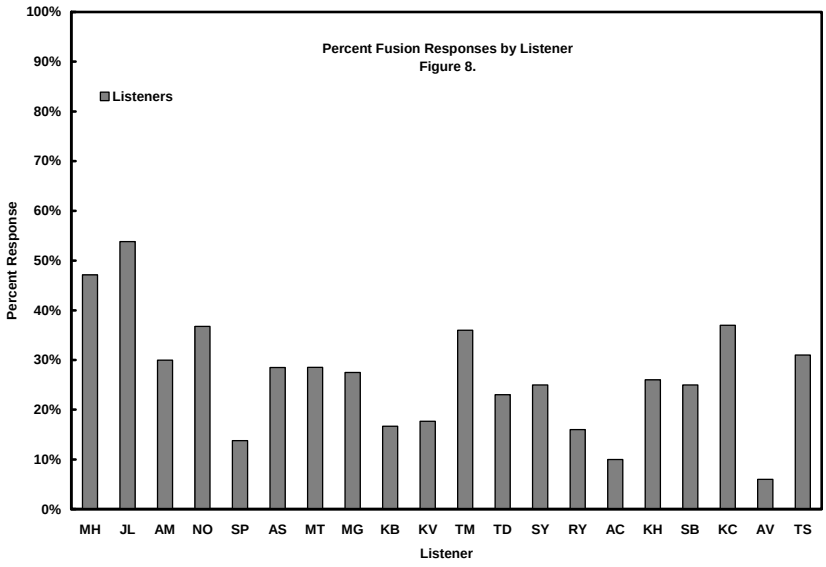

Figure 8. Percent fusion responses by listener

Figure 8 investigates fusion responses by looking at the variability across listeners. There is substantial variability in individual listeners' abilities to integrate discrepant auditory and visual stimuli. These results suggest that the variability across listeners may have impacted the overall fusion response rate. Furthermore, these results are commensurate with previous research in auditory-visual integration, which suggest that a high degree of variability in listener integration skills is partly responsible for the variability seen in the benefit that individual listeners receive from auditory-visual presentation (Grant and Seitz, 1998).

CONFUSION MATRICES

The overall results from the present study raise several questions requiring further examination. Confusion matrices were constructed in order to look at the types of perception errors for each level of degradation (2-, 4-, 6-, and 8-channel) across listeners and talkers. In general, the types of responses and errors remained fairly stable across the four degradation levels, but as expected, as percent correct scores increased confusions became less prevalent. The 2-channel matrix delivers the most interesting information for the present study.

RESPONSE

	BAT	PAT	MAT	GAT	CAT	ZAT	TAT	SAT	
BAT	**58.37%**	2.19%	8.76%	2.79%	1.79%	0.40%	0.20%	79.68%	
PAT	5.26%	**54.51%**	0.56%	0.56%	11.65%		0.75%	0.75%	
MAT	5.00%	1.50%	**72.50%**	0.50%	1.50%				
GAT	30.27%	5.69%	1.17%	**36.12%**	10.37%		0.50%		
CAT	1.76%	16.37%		2.82%	**67.08%**		4.05%	0.18%	
ZAT	12.50%	4.00%	7.00%	3.00%	2.50%	**7.50%**	2.00%	2.00%	
TAT	1.50%	31.00%		2.50%	31.50%		**15.00%**	1.50%	
SAT	10.55%	2.51%	1.01%	0.50%		1.01%	0.50%	**18.09%**	

Figure 9. Confusion matrix for 2-channel response.

As seen in Figure 9 /bæt/ was correctly perceived 50% of the time, /pæt/ with 55% accuracy, /mæt/ with 73% accuracy, /gæt/ with 36% accuracy, and /cæt/ with 67% accuracy. The most prevalent incorrect response for the stimulus /bæt/ was /mat/ which accounted for 9% of the total responses, for /pæt/ was /cæt/ at 11%, for /mæt/ was /hæt/ at 6%, for /gæt/ was /bæt/ at 30%, and for /cæt/ was /pæt/ at 16%.

These findings suggest that place of articulation information in the auditory signal was not well preserved during degrading at the 2-channel level. The confusion responses suggest that manner and voicing information were the most robust signal characteristics in the degraded auditory stimuli. This is evident by the proportion of confusion responses that did not reflect correct place of articulation, but rather some combination of manner and voicing characteristics.

The main manner cue that likely allows for confusion is the duration of turbulence in the signal, e.g., in distinguishing stops from fricatives. It may be that the replacement of spectral fine structure with noise may cause some stops to be identified as fricatives. The discrepancies in voicing are caused by timing issues, including noise duration and voice onset time. These subtle cues may be overridden by the noise structure as well.

The confusion matrices raise further questions about the stimuli employed in the present research. If the auditory signal degrading results in confusion of place of articulation, does this confusion result in “other” responses rather than integration responses? If so, then it can be assumed that the degraded information in the auditory signal does not allow listeners to integrate auditory and visual information effectively. This may suggest that the amount of information available in the auditory signal does in fact affect multimodal integration and benefit.

CHAPTER 5

SUMMARY AND CONCLUSION

The present study has examined the role of the auditory signal in auditory-visual integration. The purpose of this study was to determine if removing information from the auditory signal affects auditory-visual integration, how the process of integration is altered as the amount of available information in the auditory signal is changed and if the quality or amount of information in the auditory signal is at least partly responsible for the large degree of variability seen in the amount of benefit listeners receive from auditory-visual integration. From a clinical perspective, the present study may offer insights for signal processing strategies of cochlear implants and hearing aids with regard to providing listeners with the necessary auditory information and in the design of aural rehabilitation programs so that listeners can take full advantage of the benefits of auditory-visual integration

AUDITORY-VISUAL INTEGRATION

The present study revealed that removing information from the auditory signal negatively impacted the speech perception abilities of listeners in both the A and AV conditions.

Does a loss of information in the auditory signal inhibit auditory-visual integration? Although overall perception performance was decreased, results indicated that listeners remained able to integrate auditory and visual cues and receive benefit from this integration even though information was missing from the auditory signal. A significant effect of the level of degrading of auditory stimuli on perception performance was revealed. Listeners performed increasingly better on tasks of speech perception in both conditions when there was a greater amount of information available in the degraded auditory signal. These results suggest that listeners take advantage of the additional auditory information and use it to enhance their speech perception. Furthermore, across all levels of degrading AV performance was consistently better than A performance, suggesting that listeners are indeed able to integrate degraded auditory and visual cues.

How does the integration process change as the information available in the auditory signal is altered? Figure 2 and Figure 6 showed that auditory-visual integration remained relatively consistent across the different levels of auditory signal degrading. These results indicate that although systematically removing information from the auditory signal does impact overall speech perception performance in the A and AV conditions, listeners achieve the same amount of auditory-visual integration regardless of the amount of available information in the auditory signal. It is possible that type of auditory signal degrading employed in the present study did not remove substantially different amounts of information from the 2, 4, 6 and 8 channel stimuli to produce differences in auditory-visual integration across the levels of degradation, a finding that is similar to those noted by Grant et al. (2007). However, it is also possible the information that was removed from the auditory signal by degrading was not provided by the visual

cues and thus, relatively little new information was added and speech perception was not further enhanced.

McGurk Integration

Interestingly, results of the present study yielded a smaller than expected number of McGurk-type fusion responses across all participants. This finding was unexpected in comparison with previous studies in this laboratory, in which fusion integration rates were near 50%-60% of all discrepant responses. The rate of fusion responses for discrepant stimuli in the present study was also decreased with regards to other research using similar degraded auditory stimuli (Grant et al., 2007; Ross et al., 2006; Grant and Seitz, 1998). One key difference between the available research in auditory-visual speech perception and integration and research produced in this laboratory is the type of response task. In other laboratories, closed-set responses are typically used in discrepant AV tasks. In this laboratory we have used an open-set response task in all studies utilizing discrepant AV speech perception. The results of the present study indicate further investigation of the type of response task and resulting rates of fusion integration. Such research may have significant implications for designing aural rehabilitation training to maximize integration benefit.

Does removing information from the auditory signal affect auditory-visual integration ability for discrepant stimuli? McGurk and MacDonald (1976) found that adult listeners were able to integrate discrepant [un-degraded] auditory and visual stimuli about 90-92% of the time. The response rates in the present study, 23-32%; it is imply

that removing information from the auditory signal impacts auditory-visual integration for discrepant stimuli. This finding might also indicate that a loss of information in the auditory signal may affect the integration of different types of auditory stimuli (congruent vs. discrepant) or in different ways.

CONFUSIONS

The confusion matrices revealed some interesting data with regard to what information was retained in the degraded auditory stimulus and how the loss of certain speech cues impacts auditory-visual integration.

What information is extracted and used from the auditory signal during multimodal integration? Confusion matrices for percent correct identification revealed a loss of place of articulation information as a result of auditory signal degrading. In addition, for discrepant stimuli the largest percentages of "other" responses were not interpreted as integration responses; thus, it might be assumed that the loss of place of articulation information in the degraded auditory signal impeded listeners from integrating discrepant auditory and visual stimuli effectively. However, visual cues provide good place of articulation information and auditory-visual integration would be expected to overcome the affects of auditory signal degrading on place of articulation. What this finding might suggest, is that place of articulation is an important auditory cue for auditory-visual integration and that a loss of information available in the auditory signal does in fact affect this multimodal integration process.

TALKER VARIABILITY

Another interesting finding of the present study was the substantial amount of variability seen across talkers. Differences across talkers in the degree of benefit provided in the A and AV conditions were examined in Figure 3 and Figure 4. These results show that when a limited amount of information is available in the auditory signal there is a great deal of variability in listener performance. As the information in the auditory signal increases performance variability decreases. These results imply that when limited auditory information is available the individual characteristics of each talker have a greater effect on auditory perception. These results also suggest that the variability seen across talkers for the 2-, 4- and 6-channel stimuli may have contributed to the lack of significant differences seen in auditory-visual integration for different numbers of channels. Results of the present study suggest that an in-depth analysis of the characteristics of individual talkers as a contributing factor to the variability seen in auditory-visual integration.

CLINICAL RELEVANCE

The present study may offer insights for signal processing strategies of cochlear implants and hearing aids with regard to providing listeners with the necessary auditory information and the design of aural rehabilitation programs so that listeners can take full advantage of the benefits of auditory-visual integration.

Speech perception performance in both the A and AV conditions improves as more information is available in the auditory signal. Performance is always greater in the

AV condition when compared to A regardless of the level of auditory signal degrading, suggesting that listeners take advantage of the additional information provided in the auditory signal as well as new information provided by visual cues. It has been suggested for a long time that hearing impaired listeners who utilize hearing aids and cochlear implant devices benefit from greater amounts of auditory information during speech perception; however, there is a certain point where adding more information to the processed auditory signal does not further enhance speech perception. Results of the present study indicate that as listeners receive more auditory information they will continue to achieve the same amount of auditory-visual integration and will therefore achieve higher levels of auditory-visual speech perception as the amount of information in the auditory signal increases. So, in order for device users to take advantage of this benefit, signal processing strategies for these devices should aim at providing increased amounts of auditory signal information. Exploration of the number of channels for these devices and AV speech perception should be explored to determine how much information is necessary for listeners to achieve optimal performance.

Results of this study also indicate that regardless of the level of auditory signal degrading, listeners remained able to achieve auditory-visual integration. Thus, listeners with mild or significant degrees of hearing loss will benefit from this ability and will achieve higher levels of speech recognition when auditory and visual cues are available. Grant et al. (2007) also suggest that as greater amounts of information are removed from the auditory signal, the amount of auditory-visual integration that listeners achieve is increased and visual cues provide further enhancement of speech perception. Thus, auditory-only (A) performance may not be the best indicator of speech perception

abilities for individuals with substantial degrees of hearing loss. In AV conditions these individuals may achieve far more speech recognition than what might be predicted by A performance. These findings have two important implications, 1) listeners with even severe to profound degrees of hearing loss should always be encouraged to complete a hearing aid trial as their speech perception abilities in AV conditions may be far better than expected, 2) and auditory-visual integration should be emphasized during aural rehabilitation programs for all hearing impaired listeners, regardless of degree of hearing loss or device used.

The present study may also provide important insight with regard to the process of auditory-visual integration and implications for aural rehabilitation programs. Research suggests that individuals can benefit from auditory training and that programs which focus on training in the A condition can improve a listeners ability to detect and perceive speech using the auditory modality. There is also evidence that individuals can benefit from speechreading training to heighten this ability. Results of the present study support findings noted by Grant and Seitz (1998) suggesting that auditory-visual integration is a skill that is independent of A or V speech perception. So, it is possible that this skill should also be trained independently in order for individuals to enhance this ability and that training should be different than methods employed for auditory and speechreading training. What is less clear is how clinicians would know if they are actually training auditory-visual integration as opposed to auditory perception. One way to ensure that listeners are receiving training for this ability may be to use discrepant or McGurk-type stimuli. This would allow clinicians to determine if listener performance in auditory-visual integration improves with training, because responses for these stimuli

will reflect whether or not integration is taking place. Exploration of auditory-visual training and the methods for training this ability are warranted.

Finally, the degrading method employed in the present study was originally developed by Shannon et al. (1995) to approximate the auditory signal available to a cochlear implant device user and thus, future studies could examine the present results in light the findings in the now substantial literature evaluating auditory-visual integration in cochlear implant users.

Overall, results of the present study indicate that the information in the auditory signal plays a significant role in auditory-visual speech perception and that the perception of listeners is enhanced as more information is available in the auditory signal. Although A and AV perception performance may be decreased listeners are still able to integrate auditory and visual cues and benefit from this integration despite a loss of auditory information. However, when the auditory signal is compromised using the methods employed in the present study, differences in the amount of information available in the auditory signal do not appear to change auditory-visual integration. Since auditory-visual integration for congruent and discrepant stimuli remained stable across all levels of auditory signal degrading, no indication of "optimal integration" performance was observed. The reduced number of McGurk-type integrations may suggest that a loss of place of articulation information has a negative impact on the integration of discrepant stimuli and plays a small role in auditory-visual integration ability. The auditory signal clearly plays a role in auditory-visual speech perception and at least a minor role in integration ability. Further exploration of the listener, talker and auditory signal and their

role in the variability seen in the amount of benefit listeners receive from auditory-visual integration.

LIST OF REFERENCES

Anderson, C. (2007). *Auditory and Visual Characteristics of Individual Talkers in Multimodal Speech Perception*. The Ohio State University Department of Speech and Hearing Science. Unpublished Honors Thesis, Project Advisor: Janet M. Weisenberger, Ph.D.

Andrews, B. (2007). *Auditory and Visual Information Facilitating Speech Integration*. The Ohio State University Department of Speech and Hearing Science: Unpublished Honors Thesis. Project advisor: Janet Weisenberger, PhD.

Bellis, T. (2003). *Assessment and management of central auditory processing disorders in the educational setting: From science to practice.* Canada: Delmar Learning, Thompson Learning LLC.

Blamey, P., Cowan, R., Alcantara, J., Whitford, L& Clark, G. (1989) Speech perception studies using a multichannel electrotactile speech processor, residual hearing, and lip-reading. *Journal of the Acoustical Society of America, 85, (6),* 2593-2607

Braida, L.D. (1991). Crossmodal integration in the identification of consonant segments. *Journal of Experimental Psychology, 12,* 647-677.

Calvert, G., Campbell, R., Brammer, M. (2004). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10(11),* 649-657

Danhauer, G., Garnett, C., Edgerton, B. (1985). Older persons' performance on auditory, visual, and auditory-visual presentations. *Ear and Hearing, 6(44),* 191-197.

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *Journal of the Acoustical Society of America, 97,* 585-592.

Ernst, M.O., Banks M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415(6870),* 429-33.

Falchier, A., Clavagnier, S., Barone, P., and Kennedy, H. (2002). Anatomical Evidence of Multimodal Integration in Primate Striate Cortex. *Journal of Neuroscience, 22(13),* 5749-5759.

Grant, K.W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *Journal of the Acoustical Society of America 109(5),* 2272-2275.

Grant, K.W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective. *Journal of the Acoustical Society of America, 112 (1),* 30-33.

Grant, K.W., Braida, L.D., & Renn, R.J. (1991). Single band amplitude envelope cues as an aid to speechreading. *Quarterly Journal of Experimental Psychology, 43A,* 621-645.

Grant, K.W., Greenberg, S., Poeppel, D., van Wassenhove, V. (2004). Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing. *Seminars in Hearing 25,* 241-255.

Grant, K.W., & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America, 104,* 2438-2450.

Grant, K.W., & Seitz, P.F. (2000a.). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America, 108,* 1197-1208.

Grant, K.W., & Seitz, P.F. (2000b.). The recognition of isolated words and words in sentences: Individual variability in the use of sentence context. *Journal of the Acoustical Society of America, 107,* 1000-1011.

Grant, K.W., Walden, B.E., & Seitz, P.F. (1998). Auditory-visual speech recognition by hearing impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103,* 2677- 2690.

Grant, K.W., Tufts, J.B., & Greenburg, S. (2007). Integration efficiency for speech perception within and across sensory modalities by normal hearing and hearing impaired individuals. *Journal of the Acoustical Society of America, 121(2),* 1164-1176.

Grant, K.W., van Wassenhove, V., & Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony, *Speech Communication, (Eds.), 44(1-4),* 43-53.

Hadar, U., Steiner, T., Grant, E., & Rose, F. (1983). Head movement correlates of juncture and stress at sentence level. *Language and Speech, 26 (Pt 2),* 117-29.

Helfer, K.S. (1998). Auditory and auditory-visual recognition of clear and conversational speech by older adults. *Journal of the American Academy of Audiology, 9(3),* 234-42.

Jackson, P.L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *Volta Review, 90 (5),* 99-114.

Ladefoged, P. (1996). *Elements of acoustic phonetics*. The University of Chicago, Chicago, IL: The University of Chicago Press.

Massaro, D. (1972). Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review, (79), 2,* 124-145.

Massaro, D.W. (1987). *Speech perception by ear and eye: A paradigm for psychology inquiry.* Hillsadale, NJ: Lawrence Erlbaum.

Massaro, D.W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge, MA: MIT Press.

Massaro, D.W. & Cohen, M.M. (2000). Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception. *Journal of the Acoustical Society of America, 108,* 784-789.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

Moore, B. (2003). Psychology of Hearing. San Diego, CA: Academic Press.

Munhall, KG, Jones, JA Callan, DE, Kuratate, T. & Vatikiotis-Bateson, E (2004). Visual prosody and speech intelligibility: Head movment improves auditiory speech perception. *Psychological Science, 15(2),* 133-137.

Munhall, K.G., Kroos, C., Jozan, C., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perceptions and Psychophysics, 66,* 574 – 583.

Navarra, J., Vatakis, A., Zampini, M., Faraco, S., Humphreys, W. & Spence, C. (2005) Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research, 25(2),* 499-507.

Nicholsen, K., Baum, S., Cuddy, L., & Munhall, K., (2002). Case of impaired auditory and visual speech prosody perception after right hemisphere damage. *Neurocase, 8(4),* 314 – 322.

Ross, L., Saint-Amour, D., Javitt, V., & Foxe, J. (2006) Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17(5),* 1147-1153.

Schwartz, B. & Savariaux, Z. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition, 93(2*), B69-B78.

Shannon, R.V., Zeng, F.G., Wygonski, J., Kamath, V., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270,* 303-304.

Shannon, R.V., Zeng, F.G., & Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *Journal of the Acoustical Society of America, 104,* 2467-2475.

Spence, C., Ranson, J., & Driver, J. (2000) Cross-modal selective attention: on the difficulty of ignoring sounds at the locus of visual attention. *Perception and Psychophysics, 62(2),* 410-424.

Srinivasan, R. & Massaro, D. (2002) Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. *Language and Speech, 46(1*), 1-22.

Sumby, W.H. & Pollack, I. (1954). Visual contributions to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212-15.

Summerfield, A.Q. (1987). Some preliminaries to comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds), *Hearing by eye: The psychology of lipreading.* Hove, UK: Lawrence Erlbaum Associates Ltd.

Van Wassenhove, V., Grant, K.W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Neuroscience, 102(4),* 1181-1186.

Vatakis, A. & Spence, C. (2006). Evaluating the influence of frame rate on the temporal aspects of audiovisual speech perception. *Neuroscience Letters, 104(1-2*), 132-136.

Ver Hulst, P. (2006). *Visual and auditory factors facilitating multimodal speech integration.* The Ohio State University Department of Speech and Hearing Science, Unpublished Honors Thesis. Project advisor: Janet Weisenberger, Ph.D.

Walden, B., Busacco, D., & Montgomery, A. (1993). Benefit from visual cues in auditory-visual speech recognition by middle-aged and elderly persons. *Journal of Speech and Hearing Research,* 36, 431-436.

Wightman, F., Kistler, D., & Brungart, D. (2006). Informational masking of speech in children: Auditory-visual integration. *Journal of the Acoustical Society of America,* 119(6), 3940-3949.

Zampini, M., Shore, D.I. & Spence, C. (2003). Audiovisual temporal order judgments. *Experimental Brain Research, 152(2).*